14

USING CORPORA IN RESEARCH ON SECOND LANGUAGE PSYCHOLINGUISTICS

Sandra C. Deshors and Stefan Th. Gries

Introduction

Second language (L2) psycholinguistics has long been associated with experimental methodologies. However, over the past thirty years corpus studies have increasingly attracted scholars' attention as useful complementary or even alternative empirical approaches to address psycholinguistic questions in second language acquisition (SLA) in areas including processing, sentence complexity, fluency, lexical diversity, vocabulary acquisition, cross-linguistic transfer, and lexical knowledge in heritage language acquisition. Corpora are increasingly contributing to state-of-the-art research methodologies in the field although corpus applications emerged only relatively recently in L2 psycholinguistics and therefore often lack the methodological sophistication of existing experimental research (Gries, 2014b, p. 16). Across corpus linguistics (CL) and (L2) psycholinguistics, L2 acquisition is operationalized differently: While psycholinguistic research often focuses on comprehension (rather than production) of relatively small samples of written language and uses accuracy and reaction times as dependent variables, corpus research usually utilizes written production data. Further, many corpus-based analyses focus on the conditional probability of occurrence of a given form or meaning (as predicted by contextual information from the corpus).

The contribution of CL to SLA is most noticeable within cognitive-linguistic frameworks including constructionist, usage-based, and exemplar-based models of language acquisition and use. In line with Lakoff's Cognitive Commitment, which assumes that language and linguistic organization reflect general principles of cognition (Lakoff, 1991), these frameworks assume that (i) language acquisition, representation, and processing are largely explicable with reference to mechanisms of domain-general cognition, (ii) language use involves cognitive events that ultimately shape the linguistic system (Kemmer & Barlow, 1999), and (iii) speakers' knowledge of lexical items correlates with and their uses in grammatical contexts (e.g., Langacker, 1987; Goldberg, 2006). Corpus data have gained recognition as "a source of relevant linguistic data and consequently, quantitative and statistical tools now count as central methodologies" (Gries, 2014a, p. 280). In this chapter, we use *usage-based* as an umbrella term for the previously listed cognitive-linguistic frameworks.

Psycholinguistics and CL are intrinsically linked to the notion of *frequency*:[1] Usage frequency and repetition are central at all levels of language (Ellis, 2002) and are important for L2 learning (Ellis, 2007). In psycholinguistics, token frequencies play an important role as control variables and

 DOI: 10.4324/9781003018872-16

predictors in experimental studies, and frequency of use correlates with the entrenchment of an expression as a unit (i.e., a linguistic structure with an established cognitive routine). As Ellis et al. (2016, p. 45f) put it, "The more times we experience something, the stronger our memory for it, and the more fluently it is accessed" and

> [c]onstructions that are frequent in the input are processed more readily than rare constructions. Through experience, a learner's perceptual system becomes tuned to expect constructions according to their probability of occurrence in the input. . . . The same is true for the strength of the mappings from form to interpretation.
>
> *(Ellis et al., 2016, pp. 46–47)*

L2 acquisition and processing involve probabilistic/statistical knowledge (Ellis, 2002) and corpora provide access to many kinds of frequencies of (co-)occurrence in learner language necessary to understand how L2 knowledge is acquired and processed. Consequently, corpus-based psycholinguistic work often utilizes frequency as one predictor of acquisition (Gries, 2022) and usage-based linguists assume that distributional characteristics of linguistic elements reflect their functional (e.g., semantic, pragmatic) characteristics. Thus, their distributional characteristics (i.e., their frequencies of [co-]occurrence) reflect the similarity of their functional characteristics (Gries, 2017, p. 592). Of course,

> [l]anguage learners do not consciously count language statistics, their stream of consciousness concerns communication and understanding. The frequency tuning under consideration here is computed by the learner's system automatically during language usage. The statistics are implicitly learned and implicitly stored.
>
> *(Ellis et al., 2016, pp. 64)*

The usage-based literature posits that "[l]earning, memory and perception are all affected by frequency, recency, and context of usage" (Ellis et al., 2016, p. 46). In the context of (S)LA, research in the associative learning of cue-outcome contingencies (see Ellis, *passim*) has shown how domain-general learning mechanisms such as *entrenchment*, *productivity*, *recency*, *contingency*, *prototypicality*,

Textbox 14.1 Key terms and concepts

Psycholinguistic notion	*Brief description*	*Associated corpus method*
Entrenchment	Cognitive process by which a linguistic pattern is established as a cognitive routine	Token and type frequency of (co-)occurrence
Productivity	Cognitive process by which a linguistic pattern is extended to new cases	Type frequency of (co-)occurrence
Recency	Tendency to best remember information that is presented last	Dispersion, concordancing, structural/construction priming
Contingency	Reliability of a linguistic form as a predictor of a given interpretation	Co-occurrence (collocation, colligation, collostruction)
Prototypicality and salience[2]	Degree to which an expression is a central/ representative member of a category and stands out against other category members	Frequency, co-occurrence, contextual distinctiveness
Surprisal	The degree to which a linguistic choice is unexpected, given its context	Frequency, co-occurrence

salience, and *surprisal* along with perceptual activity all play an important role in the (S)LA process. While these notions have become central in usage-based research, corpus linguists over the past decade have actively developed statistical approaches to operationalize these notions. Textbox 14.1 presents a list of these cognitive mechanisms and their associated corpus methods. In the following, we discuss how these notions can be operationalized and captured quantitatively/statistically for L2-psycholinguistic research. We will zoom in specifically on Kim and Rah (2019), Ellis et al. (2016), and Wulff and Gries (2019). Finally, we discuss some advantages and limitations of corpus approaches to L2 psycholinguistics.

Methods and paradigms

Entrenchment

Frequency and entrenchment are strongly correlated (Gries, 2014a) in that frequency of use is said to promote entrenchment:

> Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence.
>
> *(Langacker, 1987, p. 59)*

Exploring the frequency of occurrence of linguistic elements in L2 corpora provides a way of understanding how language learners access, automatize, and process these elements (Gries & Ellis, 2015, p. 4): As linguistic elements recur, speakers' mental representations of linguistic systems are constantly being updated (see Ellis, 2002, p. 147; Halliday, 2005, p. 67; Gries, 2022, p. 51).

Importantly, entrenchment triggers the acquisition of linguistic elements at different levels of abstraction (i.e., more concrete vs. more abstract constructions), which is reminiscent of two different types of frequencies, namely token frequencies (i.e., the number of times an element is observed) vs. type frequencies (i.e., the number of different elements observed in a certain position or slot such as the number of different verbs within a prepositional dative construction). Thus, token frequency leads to the entrenchment of instances, whereas type frequency leads to the formation and entrenchment of more abstract schemas. This distinction is key to understanding, first, the richness of exemplar memories and their associations, and second, more abstract connectionist learning mechanisms. Accordingly, the relevance of token frequency grew more and more (Ellis, 2002).

In L2 acquisition, token frequencies and entrenchment are similarly relevant: Firstly, token frequency of linguistic elements in the input relates to age of acquisition (Casenhiser & Goldberg, 2005), to speed of lexical access (Schmid, 2000), to routinization, reduction (Aslin & Newport, 2012), and to category formation. For instance, exemplars with higher frequency are classified more accurately and as more typical (Ellis et al., 2016, p. 60f).

Token frequencies can be *absolute* or *relative* frequencies. Absolute frequencies refer to the number of times an element is observed (often normalized to per-million-words counts to compare results across differently large corpora and inform *context-free* entrenchment (i.e., information about the frequency of isolated linguistic elements independently of their context of occurrence). Relative frequencies, however, provide information about *contextual* entrenchment (i.e., information about the frequency/probability of elements given their linguistic or other context). Because usage-based linguists assume context is relevant for all linguistic processes, they primarily focus on relative frequencies. However, despite their perceived importance in psycholinguistics research, token frequencies alone are often less important than is assumed (Gries, 2022, pp. 52–53) and studies such as Adelman et al. (2006) and Baayen (2010) have begun to question the centrality of frequency in general or of frequency-as-repetition in particular and raise the importance of supplementing it with other predictors not discussed enough in psycholinguistics such as dispersion, association, and others (Baayen, 2010; Gries, 2022).

Productivity

Traditionally, productivity has been associated with type frequency (Bybee & Hopper, 2001; Goldberg, 2006, p. 99), which refers to "the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection or a syntactic construction specifying the relation among words" (Ellis et al., 2016, p. 52). As Ellis et al. (2016, p. 53) explain, "The more items in a certain position in a construction, the less likely the construction is associated with a particular item, and the more likely it is that a general category is formed over the items in that position." Thus, type frequencies inform categorization in L2, i.e., how learners build up constructional knowledge in the target language, in particular the connections between a given construction and the words in its slot(s). In L2 psycholinguistics, two studies drawing on both token and type frequencies, Godfroid and Uggen (2013) and Godfroid (2016), have discussed the (non-)productivity of the strong verb paradigm in contemporary German. Godfroid (2016) studied whether the observed learning was item- or system-based and justified the inclusion of a generalization posttest (to measure system-based learning) based on the notion that strong verbs form a fuzzy grammatical category. In learner corpus research (LCR), large-scale investigations of L2 knowledge have used collostructional analysis, which measures statistically the association between lexical items and grammatical constructions and requires token frequencies as well as the complete set of elements occurring in a construction's slot(s). We return to this method below and discuss how this approach can be integrated to L2 psycholinguistic studies.

Recency

Recency is the tendency to best remember information that is presented last. With corpora, recency can be explored via the linguistic contexts of exemplars with concordancing, which allows linguists to correlate frequencies of different linguistic choices with contextual information. Concordances are displays of linguistic instances of a search word in a central column together with their preceding and subsequent context (which can be defined in terms of numbers of words, numbers of characters, intonation units, etc.). Figure 14.1 shows concordance lines for particles used in phrasal verb constructions in the International Corpus of Learner English.

Across corpus-linguistic methods, concordances provide the most context and can lead to fine-grained analysis of many features on many dimensions (Gries, 2014a, p. 281). Their usefulness for L2 psycholinguistic research is that they provide all the context of a linguistic choice (to the extent it is represented in the corpus [annotation]), so one can often determine what happened in the recent past and how it might be correlated with the current investigated linguistic choice. Further, the cognitive value of concordance lines lies in the notion that memory is context-dependent (that context being of any nature, e.g., musical, linguistic); and information learned in a particular context is more readily remembered when that context is reinstated. Context-dependent memory is sensitive to incidental contextual information, it can recognize many different kinds of contextual similarities, and it can influence performance without awareness. Linguistically, context-based implicit learning has been

	A	B	C	D	E
1	FILE	LINE	PRECEDING CONTEXT	MATCH	SUBSEQUENT CONTEXT
2	_iclefr\FRUB1001.txt	6	times man needs and uses them even more. He shuts himself	out	of the hostile reality and escapes into his own inner world.
3	_iclefr\FRUB1002.txt	8	gook, incomprehensible to the general public. Let us not look	down	on "pleasure programmes" those that have an aim; for example: to p
4	_iclefr\FRUB1002.txt	6	d to fulfil a 24-hour schedule, the thorough information turns	out	to be, most of the time, diluted, repetitive and unuseful pieces of inf
5	_iclefr\FRUB1002.txt	6	nformation in the field it is dedicated to, but in order to build	up	a programme that is supposed to fulfil a 24-hour schedule, the thoro
6	_iclefr\FRUB1003.txt	7	If we jump	back	to the fifteenth century, we notice that Leonardo da Vinci already mi
7	_iclefr\FRUB1003.txt	7	ıchines as well as engines of war (so that many princes called	on	his competence as a military engineer).
8	_iclefr\FRUB1003.txt	14	All these examples point	out	that artists (that is to say people who give shape to their dreams and
9	_iclefr\FRUB1004.txt	8	ple to listen? It is quite natural that men are reluctant to give	up	their hegemony, but it is certainly no less natural, and essential, that
10	_iclefr\FRUB1005.txt	7	e on the one hand because human beings need money to live	on	for some of us at least to survive.
11	_iclefr\FRUB1005.txt	4	There had been "USA for Africa", "Band Aid". So they stood	up	for starvation but also for many illness, notably AIDS or leukaemia or
12	_iclefr\FRUB1005.txt	13	k at it, all in all, we need money to make science and to build	up	history, to give it to those who don't possess it, to survive and even t
13	_iclefr\FRUB1005.txt	15	So, to sum it	up	there are pros and cons in money, it al depends on what you do with
14	_iclefr\FRUB1006.txt	8	given the right to vote. In the USA, the same task was carried	out	by the NAWSA (National American Woman Suffrage Association).
15	_iclefr\FRUB1007.txt	5	h a shudder how the press covered the tragic events that tore	apart	Rumania a few years ago - it finally turned out that more people wer
16	_iclefr\FRUB1007.txt	5	rocess is widely used to distort reality and pass shameless lies	off	as irrefutable truths. I cannot help but remember with a shudder hov
17	_iclefr\FRUB1007.txt	5	nts that tore apart Rumania a few years ago - it finally turned	out	that more people were killed during the simultaneous U.S. interventi
18	_iclefr\FRUB1007.txt	3	ınd the war in Iraq in early 1991: we are not likely ever to find	out	the number of civilian casualties - 'no colateral damage' - who did no
19	_iclefr\FRUB1007.txt	4	n Israel. The list is endless and there is no point in following it	through	. What we should not fail to notice, however, is that censorship affec
20	_iclefr\FRUB1007.txt	6	us is somewhat more anatomic, but whoever is going to cook	up	the recipe that allows to distinguish infallibly between eroticism (wh
21	_iclefr\FRUB1008.txt	3	ut it is by no means over. It is men, not women who still carry	on	the sex war because their attitude remains hostile. Women continue
22	_iclefr\FRUB1009.txt	5	:an remember these last years, we really should take our hats	off	to feminists: they have been challenging a mentality established sinc
23	_iclefr\FRUB1009.txt	2	omen's right to a decent education. And the fight is still going	on	. Rightly? Has not feminism served its time? Has it not eventually dor
24	_iclefr\FRUB1009.txt	6	e for equality has been established, some countries still keep	on	violating it (mostly in countries where religion prevails over internati

Figure 14.1 Concordance lines for particles used in phrasal verb constructions (in the International Corpus of Learner English).

observed in areas such as homophone spelling (Smith et al., 1990), word-fragment completion (Ball et al., 2010), and picture naming (Horton, 2007).

Recent LCR provides a good illustration of how precisely linguistic contexts can be explored with concordancing, especially through annotating concordance for multiple linguistic predictors from linguistic contexts and increasingly the inclusion of psycholinguistic predictors relevant to corpus-based analyses. By applying a range of sophisticated (multifactorial/-variate) statistical techniques, researchers can assess how linguistic and cognitive predictors (jointly) correlate with aspects of learners' interlanguage. Many such studies model the probability of occurrence of a given form, which contrasts with psycholinguist studies in which often the focus is on production/comprehension accuracy and reaction times. Overall, combining comprehensive annotation and statistical analysis has helped corpus linguists better understand how notions central to psycholinguistics contribute to L2 acquisition with regards to word/sense entrenchment, the association/contingency of formal and functional elements, and matters of categorization, amongst others (Gries, 2014a, p. 287). In section 3.3 of this chapter, Corpora as primary data, we present and discuss one such example, Wulff and Gries (2019).

With their rich contextual information, concordances facilitate the exploration of recency effects, which can be manifested through (i) (structural) priming (which, in statistical terms, would be manifested as autocorrelation, a short-term effect) and (ii) dispersion (a kind of long-term recency, referring to the distribution of an element across texts, speakers, registers/genres, etc.). See for instance McDonough and Trofimovich (2009) for key priming research in L2 psycholinguistics and see Gries and Wulff (2005) for a corpus-based study of the syntactic priming of ditransitive and prepositional dative constructions in L2.

As for (i), priming refers to the fact that an occurrence of *x* increases the probability of *x* recurring beyond its (frequency-based) baseline. Priming occurs at all linguistic levels as well as non-linguistic levels (e.g., conceptual representation) and can result from implicit learning and the pattern extraction mechanisms assumed in usage-based frameworks (see, e.g., Rowland et al., 2012). While language acquisition researchers can control for priming by experimental design, this is still rare, and more comprehensive priming studies are mostly emerging in observational data (see Gries & Kootstra, 2017). Thus, corpus-based research has not only enriched and validated our understanding of priming, but has also been used for hypothesis generation (Gries & Kootstra, 2017) by providing data often ecologically more valid than that from an experimental setting. However, corpus-based studies on priming effects in L2 remain relatively rare and attention is needed to determine (i) what these effects look like in L2 and (ii) how they can be best accounted for statistically.

As for dispersion, this notion relates to general learning processes (Ambridge et al., 2006) and has been used most for lexis and response rates/reaction times in lexical decision tasks (Gries, 2019b) and the (even) distribution of words (in constructional slots) across, say, a whole corpus. For example, language users are more likely to experience constructions widely/evenly distributed in time/place. When that happens, contextual dispersion indicates that a construction is broadly conventionalized and temporal dispersion shares out recency effects (Gries, 2019b, p. 114). Given recent results (Baayen, 2010, Gries, 2019b, 2022), dispersion may well supersede frequency in its importance for (L2) learning: (More) evenly distributed exemplars can be assumed to be (more broadly) conventionalized and, therefore, to facilitate acquisition more. Generally, in psycholinguistics, dispersion is a central factor to be accounted for because it affects every kind of frequency of (co)-occurrence in a corpus.

In terms of computation, dispersion is best computed based on linguistically meaningful corpus parts (e.g., files/texts, sub-registers, registers/genres, language production modes); see Gries (2008) or Egbert et al. (2020). Callies (2013) illustrates well how dispersion over files/ speakers can account for individual-speaker variation. Disregarding dispersion in corpus analyses can lead to generalizations over parts of the corpus that may or may not be valid (e.g., speech-specific patterns may be attributed to written language), which can undermine all conclusions of an analysis.

Contingency

Associations (i.e., contingency) between linguistic forms and/or between them and their functions also play an essential role in all aspects of language. Based on Siyanova-Chanturia and Pellicer-Sanchez (2018) and Wray (2002), formulaic units are psychologically real and according to Durrant (2014) and Durrant & Schmitt (2009), collocational knowledge is an important aspect of learner language. For Ellis (2006, p. 7), "[l]anguage learning can be viewed as a statistical process requiring the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations." As speakers learn their L2, they acquire the ability to map forms and functions reliably by keeping track of a wide range of co-occurrence information of both their language comprehension and production (Gries & Ellis, 2015, p. 21). Contingency therefore drives associative learning (Ellis, 2016, p. 62); for instance, collocational and phraseological knowledge is central to the attainment of native-like fluency and native-like idiomaticity (Pawley & Syder, 1983).

To date, much corpus-based research on contingency has been conducted based on linguistic co-occurrences such as collocations, colligations, and collostructions (i.e., lexico-grammatical co-occurrences; see Gries & Durrant, 2020 for an overview). Specifically, contingency, as manifested by co-occurrence counts, helps to explore *what-if* relations, i.e., *what happens if the context is like this?* Corpus-based contingency work assumes that (i) "human learning is . . . perfectly calibrated with normative statistical measures of contingency" (Ellis, 2006, p. 7) and (ii) statistical associations between linguistic elements found in corpus data reflect the psychological associations in the minds of language learners (Stefanowitsch, 2006).

Therefore, many association measures (AMs) have been developed including conditional probability, (logged) $P_{\text{Fisher-Yates exact}}$, *t, z, odds ratio*, *MI* but there is currently no consensus on how to best measure contingency with regard to symmetry of association, type of metric, and frequency information. However, "different AMs offer different and complementary perspectives on collocation learning" (Gries & Durrant, 2020; see Durrant (2014) on the relationship between L2 learners' knowledge of English collocations and various measures of collocation frequency and association). One approach widely adopted in SLA corpus-based research is collostructional analysis (CA; Gries & Stefanowitsch, 2004), a family of approaches to quantify (i) degrees of attraction/repulsion of words (typically verbs) to syntactically defined words in a construction (collexeme analysis), (ii) which words are attracted/repelled by one of several constructions (distinctive collexeme analysis), and (iii) identify (dis)preferred pairs in two slots of one construction (co-varying collexeme analysis). Findings of such work inform many studies on issues such as item-based learning or generalization and the question of which words are particularly attracted to particular constructions and, therefore, likely to function as path-breaking words in constructional acquisition.

Prototypicality and salience

Concordancing and contingency are also useful to investigate the interrelated notions of prototypicality and salience. According to the influential weighted-attribute approach, a prototype is an abstract entity—not a concrete exemplar—which combines the most salient attributes of the category, wherein (i) those attributes are those with a high cue validity for the category, and (ii) the cue validity of an attribute A (e.g., flying) of object X (e.g., a sparrow) with regard to a category C (e.g., birds) is the conditional probability of X being a member of category C given that X exhibits A $p(C|A)$ (see Taylor, 2011: Section 5.2; Ellis et al., 2016).

Prototypes, or more precisely, entities close to the abstract prototypes, exhibit a variety of effects, many of which are measurable with corpora: They are acquired earlier, produced more often (i.e., they are often more frequent and more associated to certain contexts), recognized faster, invite generalizations more than more marginal category members, are perceptually more salient, etc. (Taylor, 2011). But corpora can not only help identify specific characteristics that are a part of a prototype, but they can also determine to what degree these characteristics contribute to a prototype. For example, Ellis et al. (2016) used corpora to study semantic prototypicality in L2 verb argument constructions (VACs) and build semantic networks based on verb type frequencies as extracted from VAC distributions. Their data show how quantitative measurements of semantic relations between verb types and VAC frames can be used to explore L2 speakers' linguistic knowledge based on co-occurrence patterns in corpora. Similarly, much of the work on collostructional analysis identifies the verbs most strongly attracted to certain constructions because these verbs reflect the prototypical sense(s) of a construction (e.g., *give* and *tell* for the ditransitive construction; see Gries & Stefanowitsch, 2004).

Surprisal

Surprisal is somewhat different from the other notions. On the one hand, it is a driving force of the language learning process (Ellis et al., 2016). According to Jágrová et al. (2019, p. 244), "[i]ntuitively, it can be thought of as measuring the information content conveyed by a linguistic unit [given its (preceding) context] and it appears to scale the cognitive effort required to process this information." Thus, like most usage-based notions, surprisal implies a probabilistic approach to language. For example, when a hearer hears a certain verb and, from that, expects (or predicts) a certain complementation pattern to follow, which then does not happen, the learning process is enhanced: "One consequence is that, when prediction goes wrong, it is surprisal that maximally drives learning from a single trial. Otherwise, the regularities of the usual course of our experiences add up little by little, trial after trial, to drive our expectations" (Ellis, 2016, p. 344). Within the visual world eye-tracking paradigm, this type of approach aligns with Jackson and Hopp's (2020) exploration of whether prediction failures during real-time processing drive language acquisition.

On the other hand, surprisal is also central as a moderator for some of the other notions discussed earlier. For instance, Jaeger and Snider (2008) showed that surprisal can amplify priming effects: Unexpected uses prime more than expected ones and, arguably, surprisal will also amplify salience simply because an expected expression will "stick out more" as it is processed. However, corpus studies involving surprisal in L2 acquisition research are extremely rare. One exception is Wulff et al. (2018), a corpus analysis of optional *that* complementation in native and learner English. They use surprisal—how surprising the transition is from a matrix clause to a complement clause—as a predictor in a multifactorial analysis alongside twelve other linguistic factors and find that speakers smooth over more surprising local transitions by using *that* while low-surprisal transitions (e.g., *the* as part of the complement clause subject after *I think*) feature *that* much less often.

Example studies

Corpora can play different roles and serve different purposes in L2 psycholinguistic research: They can (i) serve as methodological tools in setting up experimental studies; (ii) complement experiments (i.e., when experimental and corpus approaches are triangulated); and (iii) serve as the main source of empirical data. In this section, we illustrate each role by reviewing individual studies.

Corpora and experimental design

Amongst L2 experimentalists, native-language corpora have become widespread tools in the development of experiments: They have been used to extract word frequencies in native language (L1) to be used as predictors or control variables to assess learners' performances (Gries & Ellis, 2015) and they have allowed scholars to establish native-like baselines against which to contrast L2. This approach has proved popular in processing (e.g., Spinner et al., 2017), morphology (e.g., Matusevych et al., 2018), syntax (e.g., Hopp, 2017), collocational knowledge (e.g., Toomer & Elgort, 2019), linguistic contexts and their effects on phonolexical processing (e.g., Chrabaszcz & Gor, 2014), and constructional knowledge (e.g., Kim & Rah, 2019). In the case of Kim and Rah (2019), following Johnson and Goldberg (2013), the authors used the Corpus of Contemporary American English (COCA) to select verbs for an experiment designed to explore L2 learners' sensitivity to constructional information and learners' efficiency in integrating information from a verb and a construction in real-time processing. Corpus data were integrated into the experiment by calculating lexical verb frequencies from the COCA and by conducting a collostructional analysis to quantify how much individual lexical verbs co-occur with the investigated constructions. This helped the authors to choose experimental stimuli based on both frequencies and association strengths for lower-frequency target verbs and their higher-frequency counterparts, which ultimately increased the generalizability of the analysis. More theoretically, the authors showed that learners integrate argument roles between a verb and a construction faster when focusing on constructions rather than lexical verbs, stressing the importance of constructional information for L2 sentence processing.

Triangulating corpus-based and experimental approaches

Corpus data in L2 psycholinguistic research can involve triangulating corpus and experimental methodologies complementarily. With corpora, L2 phenomena can be often explored at a (potentially) much larger scale than experiments would allow and with greater ecological validity—with experiments, much control can be exercised but potentially at the cost of ecological validity. Ellis et al. (2016) adopted both corpus and experimental methodologies to explore the extent to which native speakers' knowledge of VACs differs from that of L2 learners and whether the type of learners' L1 in terms of verb semantics is a bias towards their knowledge of the L2. They first focused on native data and extracted and analyzed VACs from the British National Corpus to determine their contingencies with verbs. Then, for each VAC, they identified and quantified core meanings and construction semantic networks around the notions of prototypicality, semantic cohesion, and polysemy.

In a second step, they compared the VAC uses in the native data with those of German/Czech/Spanish L2 learners, which they elicited experimentally with a survey. The corpus findings established that VACs promote learning in that (i) they are Zipfian (i.e., many words are very rare, very few words are very frequent) in their verb type-token ratio constituency in usage, (ii) constructions prefer certain verbs in them, and (iii) they are coherent in their semantics. For the experimental part, native and L2 participants were administered an online VAC survey involving a free association task. The verb responses collected for each VAC were lemmatized by verb type and ordered by verb token frequencies. The authors then compared lists based on the learner responses with lists based on native English responses, and focus was given to the potential effects of frequency, contingency, and semantic prototypicality. Verb frequency in the VAC, VAC-verb contingency, and verb prototypicality co-determined learners' responses to VAC prompts, leading to the conclusion that "L2 VAC processing involves rich associations, tuned by L2 verb type and token frequencies and their contingencies of usage, which interface syntax, lexis and semantics."

Corpora as primary data

Finally, corpora can be explored as primary sources of data. As mentioned earlier, frequency, recency, and context affect SLA jointly and their mutual interaction affects how learners acquire their L2: "The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization" (Ellis et al., 2016, p. 46). LCR scholars are now exploring such interactions in corpora and how they contribute to the process of L2 acquisition and use with multifactorial/multivariate approaches involving many linguistic predictors derived from concordance lines (e.g., animacy, verb semantics, verb type, voice) and cognitive predictors operationalizing many of the above notions.

For example, Wulff and Gries (2019) targeted the syntactic alternation between verb-particle-object (VPO) vs. verb-object-particle (VOP) constructions across native English and over a dozen interlanguage varieties with a multifactorial analysis of approximately 5,000 occurrences of the two constructions from various L1 and L2 corpora. They explored how learners' syntactic choices were influenced by processing demands, input effects, and L1 typology by analyzing the joint effects of 17 predictors, including, for instance, the order of constituents in the verb particle construction (VPC), lengths of the particle and the direct object noun phrase, the complexity of the direct object, and rhythmic alternation in the VPC. They used the MuPDAR approach (Multifactorial Prediction and Deviation Analysis using Regression; see Gries & Deshors, 2014), a two-step regression procedure that computes for every learner choice what an L1 speaker would have chosen in the same context and then explores where and why the learner choices differ from the L1 speaker choices. Among other things, Wulff and Gries found that, firstly, learners overuse VPO constructions across nearly all contextual conditions. Secondly, particle stranding in the VOP construction incurs a cognitive load, which impacts learners' constructional choices more strongly in speech than in writing, presumably due to the demands of online processing in speech.

Advantages and limitations of corpus data for L2 psycholinguistics

Advantages

Corpora have much to offer to L2 psycholinguists: They offer methodological options avoiding potential problems associated with experimental designs, such as low ecological validity and input misrepresentation.

Regarding the former, by nature, experimental studies are conducted under highly controlled conditions based on carefully elicited or selected data. In contrast, as an authentic/spontaneous and highly contextualized type of data, corpora can lead to studies with a higher level of ecological validity because the data are not produced in, say, read-one-sentence-at-a-time kinds of situations.

Regarding input distribution, the controlled (i.e., well balanced) design of experiments often means that participants are exposed to unrepresentative distributions of investigated linguistic elements (Gries, 2019a), which can be problematic given that "learning effects can be observed even over a relatively small number of experimental stimuli" (Gries, 2019a, p. 235; see also Baayen et al., 2018). Put differently, there can be within-experiment learning effects given the unrepresentative input, but there can also be other within-experiment factors such as fatigue or habituation that can distort the results of experiments especially with learners whose probabilistic knowledge is not yet as robust as that of native speakers—corpora do not come with these problems.

A final advantage is that corpora allow scholars to complement experimental L2 psycholinguistic research. For example, the study of how priming effects differ across different prime-target distances in a setting not perturbed by unrepresentative experimental input can benefit much from corpus studies, which are also well suited for exploratory investigations and hypothesis-generating work.

Limitations

Despite, or in fact because of, their advantages, corpora also come with challenges: Ecological validity of data also means such data are often noisy and heterogeneous. For instance, given that conditions of language production are not always strictly controlled, linguistic contexts may vary across the uses of a given linguistic element across speakers. Further, data distribution can be problematic because data are often unbalanced/Zipfian with frequencies of words that decrease as a power function of their rank in the frequency table (Ellis et al., 2016), which can require very complex statistical analyses. Also, corpus data can present challenges in terms of (i) collinearity between predictors and (ii) the often necessary inclusion of many control variables to control statistically for what, with corpora, cannot be controlled for by (experimental) design.

Finally, metadata regarding speakers are often lacking, making it hard to fully account for the learning/sociodemographic background of the speakers in the corpus and how they may affect learners' acquisition of the L2.

Innovations and future directions

Much work remains to be done for corpus linguists involved in L2 psycholinguistic research; Textbox 14.2 summarizes the main direction for the research areas we've discussed in this chapter.

Regarding corpus compilation, various aspects of this process are important to address. Specifically, we need (more) corpora that

- are bigger (to have more data points for proper statistical modeling);
- have richer metadata (e.g., degree of motivation, attitude towards the L2, cognitive variables, etc., to know more about the speakers represented);
- are more varied—L1s, register, mode, etc. (to have more diverse data to generalize from);
- are longitudinal (to track development better than with cross-sectional studies);
- are not only accessible on a website (to compute statistics that web interfaces do not provide).

Quantitative analyses of such corpora will bring a quality to both psycholinguistic and corpus-based analyses that is as yet difficult to attain.

Regarding statistical approaches to corpora, we need more sophisticated conceptualizations of the various forms in which frequencies in corpora can operationalize cognitive/psycholinguistic notions. To return to an example from above, the field needs to recognize that frequency must be complemented with dispersion information (see again Adelman et al., 2006; Baayen, 2010, Gries, 2019b), given the evidence that frequency is a convenient, but imprecise, proxy for entrenchment: Words of extremely similar frequencies can differ massively in their dispersion and their likelihood of being known by learners. Thus, studies relying on frequency (as a predictor or as a control) alone are likely to fail in properly capturing "entrenchment" or "exposure." One can only wonder why there are dozens of measures of association, lexical diversity, lexical dispersion, but an experimentally supported variable such as dispersion is ignored.

Finally, much like surprisal, the notion of salience (and its role in SLA) remains to be operationalized in a valid corpus-based way (Gass et al., 2018). We are currently not aware of work incorporating salience in a systematic corpus-based/-driven fashion. It is conceivable that salience could in fact be considered as a central component of surprisal and, thus, be operationalized the same way (see Gries, 2019b, p. 65) but this awaits corpus-linguistic exploration.

It is our view that (methodological) innovation in L2 psycholinguistics is dependent upon close collaboration between L2 psycholinguists and corpus linguists (see Rebuschat et al., 2017 for a special issue on multidisciplinary research across experimental, computational, and corpus-based methodological boundaries in [S]LA). Both types of researchers use corpora, but to varying degrees and for different purposes: While most L2 psycholinguistics research published in flagship SLA journals does not yet offer analyses of corpora as primary data sources, learner corpus scholars do utilize corpora primarily as main data sources but they are yet to conduct analyses that fully account for comprehensive ranges of psycholinguistic predictors. By working together, L2 psycholinguists and corpus linguists will undoubtedly manage to shed more light on what it means to acquire a second language.

Textbox 14.2 Open questions and issues

Corpus compilation: (Continue to) develop larger corpora with more varied data. Continue development of longitudinal corpora.

Corpus metadata: Compile metadata databases on individuals, cognitive variables, aptitude, motivation, etc.

Statistical approaches to corpora: (Continue to) develop corpus-based ways to handle noisy data.

Salience in L2: Follow the footsteps of Wulff et al. (2018) by developing studies that operationalize and measure salience.

Notes

1 See Christiansen and Chater (2016) for a summary of the central role of frequency in psycholinguistic work. Frequency effects are often observed in reaction times: Less frequent items incur higher reaction times compared to more frequent items; note that this does not *prove* that frequency is the cause.

2 We include *salience* in Textbox 14.1 for a complete picture of the notions involved in L2 acquisition and their operationalization in corpus studies. However, due to length constraints, salience is not extensively discussed.

Further reading

Crossley, S., Kristopher, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal*, *100*(3), 702–715.

This corpus-based longitudinal study focuses on salience and the extent to which L2 learners are more likely to produce lexical items in their L2 that are more salient in the L1 input.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Language Learning Monograph.

This volume explores verb argument constructions in first and second language acquisition, processing, and use from a usage-based theoretical perspective. While reaffirming the value of interdisciplinarity, the authors show us the power of corpus data to better understand L2 acquisition and processing.

Gries, St. Th. (2019). *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Brill (Distinguished Lectures in Cognitive Linguistics series).

This volume connects psycholinguistics and CL by focusing on the operationalization and measurement of cognitive notions such as frequency, dispersion, and context for quantitative analysis.

References

Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, nor word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Ambridge, B., Theakston, A., Lieven, E., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2), 174–193.

Aslin, R., & Newport, E. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*, 170–176.

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

Baayen, H., van Rija, J., de Catc, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In *Mixed-effects regression models in linguistics* (pp. 49–69). Springer.

Ball, L., Shocker, J., & Miles, J. (2010). Odour-based context reinstatement effects with indirect measures of memory: The curious case of rosemary. *British Journal of Psychology*, *101*(4), 655–678.

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. John Benjamins.

Callies, M. (2013). Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics*, *18*(3), 357–390.

Casenhiser, D., & Goldberg, A. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, *8*, 500–508.

Chrabaszcz, A., & Gor, K. (2014). Context effects in the processing of phonolexical ambiguity in L2. *Language Learning*, 415–455.

Christiansen, M., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition and processing*. The MIT Press.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations. *International Journal of Corpus Linguistics*, *19*(4), 443–477.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, *47*(2), 157–177.

Egbert, J., Burch, B., & Biber, D. (2020). Lexical dispersion and corpus design. *International Journal of Corpus Linguistics*, *25*(1), 89–115.

Ellis, N. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

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

Ellis, N. (2007). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–27.

Ellis, N. (2016). Salience, cognition, language complexity, and complex adaptive systems. *Studies in Second Language Acquisition*, *38*(2), 341–351.

Ellis, N., Römer, U., & O'Donnell, M. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Language Learning Monograph Series. Wiley-Blackwell.

Gass, S., Spinner, P., & Behney, J. (Eds.). (2018). *Salience in second language acquisition*. Routledge.

Godfroid, A. (2016). The effects of implicit instruction on implicit and explicit knowledge development. *Studies in Second Language Acquisition*, *38*(2), 177–215.

Godfroid, A., & Uggen, S. (2013). Attention to irregular verbs by beginning learners of German—an eye movement study. *Studies in Second Language Acquisition*, *35*(2), 291–322.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Gries, St. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437.

Gries, St. Th. (2014a). Corpus and quantitative methods. In J. Taylor & J. Littlemore (Eds.), *The bloomsbury companion to cognitive linguistics* (pp. 279–300). Bloomsbury.

Gries, St. Th. (2014b). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: Some necessary clarifications. In N. Gisborne & W. Hollmann (Eds.), *Theory and data in cognitive linguistics* (pp. 15–48). John Benjamins.

Gries, St. Th. (2017). Corpus approaches. In B. Dancygier (Ed.), *Cambridge handbook of cognitive linguistics* (pp. 590–606). Cambridge University Press.

Gries, St. Th. (2019a). Priming of syntactic alternations by learners of English: An analysis of sentence-completion and collostructional results. In J. Egbert & P. Baker (Eds.), *Using corpus methods to triangulate linguistic analysis* (pp. 219–238). Routledge.

Gries, St. Th. (2019b). *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Distinguished Lectures in Cognitive Linguistics Series. Brill.

Gries, St. Th. (2022). On, or against?, (just) frequency. In H. Boas (Ed.), *Directions for pedagogical Construction Grammar: Learning and teaching (with) constructions* (pp. 47–72). De Gruyter Mouton.

Gries, St. Th., & Deshors, S. (2014). Using regressions to explore deviations between interlanguage and native language: Two suggestions. *Corpora*, *9*(1), 109–136.

Gries, St. Th., & Durrant, P. (2020). Analyzing co-occurrence data. In M. Paquot & St. Th. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 141–159). Springer.

Gries, St. Th., & Ellis, N. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 1–28.

Gries, St. Th., & Kootstra, G. (2017). Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, *20*(2), 235–250.

Gries, St. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on "alternations". *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Gries, St. Th., & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, *3*, 182–200.

Halliday, M. (2005). *Computational and quantitative studies*. Continuum.

Hopp, H. (2017). Cross-linguistic lexical and syntactic co-activation in L2 sentence processing. *Linguistic Approaches to Bilingualism*, *7*, 96–130.

Horton, W. (2007). The influence of partner-specific memory associations on language production: Evidence form picture-naming. *Language and Cognitive Processes*, *22*(7), 1114–1139.

Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism*, *24*(5-6), 895–911.

Jaeger, F., & Snider, E. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the 29th annual Cognitive Science Society* (pp. 1061–1066).

Jágrová, K., Avgustinova, T., Stenger, I., & Fischer, A. (2019). Language models, surprisal and fantasy in Slavic intercomprehension. *Computer Speech and Language*, *53*, 242–275.

Johnson, M., & Goldberg, A. (2013). Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, *28*, 1439–1452.

Kemmer, S., & Barlow, M. (1999). Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 7–28). Center for the Study of Language and Information.

Kim, H., & Rah, Y. (2019). Constructional processing in a second language: The role of constructional knowledge in verb-construction integration. *Language Learning*, *69*, 1022–1056.

Lakoff, G. (1991). Cognitive versus generative linguistics: How commitments influence results. *Language and Communication*, *11*(1–2), 53–62.

Langacker, R. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.

Matusevych, Y., Alishahi, A., & Backus, A. (2018). Quantifying cross-linguistic influence with a computational model: A study of case-marking comprehension. *Linguistic Approaches to Bilingualism*, *8*, 561–605.

McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. Routledge.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). Longman.

Rebuschat, P., Meurers, D., & McEnery, T. (Eds.). (2017). Language learning research at the intersection of experimental, computational, and corpus-based approaches. Special issue of *Language Learning*, *67*.

Rowland, C., Chang, F., Ambridge, B., Pine, J., & Lieven, E. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition*, *125*(1), 49–63.

Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). (2018). *Understanding formulaic language: A second language acquisition perspective*. Routledge.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. de Gruyter.

Smith, S., Heath, F., & Vela, E. (1990). Environmental context-dependent homophone spelling. *The American Journal of Psychology*, *103*(2), 229–242.

Spinner, P., Foote, R., & Upor, R. (2017). Gender and number processing in second language Swahili. *Linguistic Approaches to Bilingualism*, *8*, 446–476.

Stefanowitsch, A. (2006). Distinctive collexeme analysis and diachrony: A comment. *Corpus Linguistics and Linguistic Theory*, *2*(2), 257–262.

Taylor, J. R. (2011). Prototype theory. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics* (pp. 643–664). De Gruyter Mouton.

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*, 405–439.

Wray, A. (2002). *Formulaic Language and the lexicon*. Cambridge University Press.

Wulff, S., & Gries, St. Th. (2019). Particle placement in learner English: Measuring effects of context, first language, and individual variation. *Language Learning*, *69*(4), 873–910.

Wulff, S., Gries, St. Th., & Lester, N. (2018). Optional that in complementation by German and Spanish learners. In A. Tyler, L. Huan, & H. Jan (Eds.), *What is applied cognitive linguistics? Answers from current SLA research* (pp. 99–120). De Gruyter Mouton.