This is an Accepted Manuscript of a book chapter published by Routledge/CRC Press in:
Lozano, C. (2021). Generative approaches. In N. Tracy-Ventura & M. Paquot (Eds.), The Routledge Handbook of Second Language Acquisition and Corpora (pp. 213-227). Routledge.
https://doi.org/10.4324/9781351137904

16

Generative approaches

Cristóbal Lozano

Abstract

This chapter discusses learner corpora in generative approaches to second language acquisition (GenSLA) research. It first reviews the core issues in GenSLA over the past decades and then discusses the kinds of research methods and data that have been traditionally used in GenSLA, their advantages and limitations, and how learner corpus methods can contribute to hypothesis-testing (Gen)SLA approaches. The chapter finally illustrates such a GenSLA-based, learner corpus research (LCR) approach by reviewing studies on tense-aspect marking and anaphora resolution. It concludes with future directions on how learner corpora can better inform hypothesis-testing GenSLA via the triangulation of experimental and corpus data.

Introduction

GenSLA is an explanatory, theory-driven approach that focuses on the acquisition of linguistic knowledge in an L2, leaving aside individual factors and social aspects. Following Chomsky's (1986) three major questions (*What constitutes knowledge of language? How is knowledge of language acquired? How is knowledge of language put to use?*), GenSLA has focused more on L2 competence (how the implicit or unconscious knowledge of the L2 is acquired and represented in the mind) than on performance (how such knowledge is put to use), and on how Universal Grammar (UG) shapes such knowledge (Hawkins, 2001, 2007; Rothman & Slabakova, 2018; White, 1989, 2003, 2012). L2 learner's developing grammatical knowledge, standardly known as interlanguage grammars (ILGs), are the result of multiple factors: (i) influence from learners' L1 (transfer), (ii) L2 input, (iii) universal and cognitive mechanisms common to all learners irrespective of their L1.

UG has been reconceptualised over the decades. Under the Principles and Parameters (P&P) theory (Chomsky, 1981), UG consists of innate universal principles common to all languages and language-specific parameters of variation that must be fixed by experience (input). For example, all languages have subjects, but the null-subject parameter stipulates that overt pronominal subjects must be realised in some languages (English: *He found a solution*) but null pronouns (*Ø*) are licensed in others (Spanish: *Él/Ø encontró una solución*). Parameters were envisaged as clusters of properties, e.g., null-subject languages additionally allow Subject-Verb inversion (*Existen los androides*) whereas non null-subject languages do not (**Exist androids*). In L1 acquisition, the input the child is exposed to (English or Spanish) is simply a trigger to

set the parameter. In L2 acquisition, the learner must reset parameters when the L1-L2 parametric values do not coincide (White, 1989).

After UG, the Minimalist Program (MP) (Chomsky, 1995) became the general framework for the study of generative approaches to language acquisition. The minimalist approach capitalises on the lexicon: variation between languages is located in the features of functional categories. UG consists of a computational system and a lexicon containing a universal inventory of features, a set of which are selected in the process of L1 acquisition. For example, the functional category Aspect, which may be realised in different morphological ways across languages, can host a variety of features ([perfective], [imperfective: continuous], [imperfective: habitual], [imperfective: progressive]). Differences between languages therefore lie in which features (out of the inventory of UG) are selected in each language, and how these features are assembled onto functional categories. The GenSLA question is whether learners can acquire the L2 features that were not selected during L1 acquisition, and, if they were selected, whether they can remap them onto new L2 forms. After the publication of the MP, GenSLA saw an upsurge of different feature-based theories trying to account for the (lack of) success in L2 acquisition, like the Failed Functional Features Hypothesis, the Feature Reassembly Hypothesis, the Missing Surface Inflection Hypothesis, the Interpretability Hypothesis and others (for an overview of those theories see White, 2012). The notion of ultimate attainment is relevant in these models: whereas children eventually attain native-like competence in their L1, learners fail to do so in their L2 as a result from either deficits in knowledge (learners' inability to fully acquire new functional features in their L2) or processing limitations (their inability to (re)map the features onto their corresponding morphological exponents), sometimes as a result of a critical period after which new features are unacquirable (Liceras, Zobl & Goodluck, 2008 and articles therein for overviews). Such deficits typically result in learners' incomplete and divergent grammars when compared to native grammars (Sorace, 2000).

Recent developments (Chomsky, 2000, 2005, 2011) within the 'biolinguistic' approach focus on the properties which (external) interface conditions impose on the design of the language faculty (UG). In GenSLA, attention has shifted from UG access, parameter (re)setting and feature reassembly to how the computational system of UG (i.e., the syntactic module) interfaces with other language-internal modules (lexicon) and language-external modules (phonology, semantics/pragmatics). Much of recent GenSLA research has been motivated by the syntax-discourse and syntax-semantics interfaces (Montrul, 2011; Rothman & Slabakova, 2011; White, 2011), which have been argued to be a locus of residual variability and optionality in ILGs (Sorace, 2000, 2011).

Core issues and topics

L1 transfer and learner corpora

The **role of L1 transfer** in L2 acquisition, which is a key issue in SLA, has been reinterpreted over the years in GenSLA. The Principles & Parameters (P&P) model predicted learners to transfer their L1 parametric settings (and their associated cluster of properties) onto their L2. Feature-based models like the Feature Reassembly Hypothesis (FRH) (Lardiere, 2009) envisage SLA as a feature (re)assembly process. Learners must (i) acquire new L2 features if they are not present in their L1, and, (ii) if those features are present in both L1 and L2, abandon their L1 feature-form mapping and reassemble the features onto the corresponding L2 forms. Theoretically-informed approaches like the FRH allow researchers to finely discriminate

between L1 transfer vs universal effects, so corpus data can be used to investigate such hypotheses in a more nuanced way, as will be illustrated below in the subsection on Tense-Aspect marking. Many LCR studies have traditionally been more descriptive than explanatory (Myles, 2015) and have readily attributed learners' errors to L1 transfer (Callies, 2015, p. 49), when in fact many are developmental or universal in nature. A theoretically-informed approach to L2 corpus data, such as one that examines the predictions of the FRH, therefore provides a more nunanced understanding of L1 effects (if any).

Frequency, input and access to UG

Frequency effects have been treated cautiosly in GenSLA, since it is well known that highly frequent functional categories (e.g., third person singular *–s*, past-tense *–ed*) and even the most frequent category in English (the definite article *the*) are not acquired early in L2 English and typically pose persistent problems (Hawkins, 2001; Lardiere, 2007). GenSLA provides theoretically-motivated accounts why some properties and not others are persistently problematic despite their high frequency in the input. The input often underdetermines the complex and subtle linguistic knowledge the learner eventually attains. This is known as the poverty of the stimulus (**PoS**) and is a key issue in generative L1 acquisition: the child eventually knows linguistically more than what is obvious from the input thanks to UG (Chomsky, 1986). A classic question in GenSLA, known as the access-to-UG debate, is whether this innate device (UG) is still available to the adult learner in L2 acquisition. L2 knowledge of PoS phenomena is a crucial piece of evidence to argue for access to UG in SLA. For example, the the Overt Pronoun Constraint (Montalbetti, 1986) is a universal stating that a null pronominal subject (*Ø*) in a subordinate clause can co-refer with a quantified subject antecedent

in a main clause but an overt pronominal subject cannot as it must refer to someone else (*Nadie$_i$ dice que Ø$_i$/él$_{j/*i}$ ha aprobado el examen* 'Nobody$_i$ says that Ø$_i$/he$_{j/*i}$ has passed the exam'). Experimental data (Lozano, 2003) show that English learners of L2 Spanish have native-like knowledge of the OPC from early stages, even though their knowledge cannot derive from (i) their L1 English (only overt pronominal subjects are allowed in English), (ii) language teaching or textooks (as the OPC is never explained), and (iii) the Spanish input, since the structural combination *[quantifier$_i$ ... [Ø$_i$]]* was not present in large Spanish native corpora (which may explain why corpora have not been used as a priviledged source of data in GenSLA).

The role of the interfaces

Interfaces have received much attention in recent GenSLA. The Interface Hypothesis, IH (Sorace, 2011, 2012) postulates that constructions that are constrained by narrow syntax (e.g., the licensing of null pronominal subjects in null-subject languages) are typically acquired early in development, whereas constructions constrained by language-external interfaces (e.g., the use of null pronouns when regulated by topic and focus at the syntax-discourse, cf. subsection on pronouns and anaphora below) are persistently problematic and lead to optionality even in near-native levels of L2 competence. This results from the complexity of simultaneously integrating and processing syntactic with discursive knowledge (but see White 2011 for an argument that not all interface properties are equally problematic). Taking the IH and White's criticism as a departure point, Lozano (2016, 2018) proposed the Pragmatic Principles Violation Hypothesis (PPVH). Not all pronominal deficits at the syntax-discourse interface are equally problematic: (i) learners often violate the pragmatic Principle of Economy since in topic-continuity contexts they produce redundant overt pronouns, which happens to be more frequent with two than with one potential antecedent; (ii) they tend to avoid ambiguity (i.e., the use of

null pronouns in topic-shift scenarios) as it would lead to a communicative breakdown therefore violating the Principle of Clarity. In short, there are pragmatic reasons why learners are more redundant and overexplicit than ambiguous, a phenomenon commonly reported in the L2 literature.

To summarise, the centrality of linguistic theory in GenSLA has two consequences for SLA: (i) a high level of prediction and explanatory power, and (ii) a way to make sense of bodies of data (Rothman & Slabakova, 2018). Both consequences are relevant for LCR. First, a high level of prediction allows the researcher to search the corpus for a particular linguistic phenomenon amongst a mass of data (hypothesis-testing approach). Second, new meaningful factors may be uncovered in that mass of data in light of the theory (hypothesis-finding approach). GenSLA corpus-based approaches should therefore make use of a top-down approach (departing from a hypothesis to interrogate the corpus) but also the bottom-up approach (exploring the corpus to find hypotheses) that has been typical in LCR (Callies, 2015; Mendikoetxea, 2014; Myles, 2007).

Main research methods

In SLA research, there is always a tension between the need for natural data vs degree of control of the data (Table 1), with some researchers favouring corpus data while others favouring experimental data. To test specific hypotheses, GenSLA researchers have traditionally favoured more controlled but less natural data (experiments) over corpus methods, which are less

controlled but more natural (Lozano & Mendikoetxea, 2013; Myles, 2007). Experiments are typically divided into offline vs online. Offline methods measure the learner's competence/knowledge of the L2 after the stimulus has been presented, whereas online methods measure performance/processing in real time as the stimulus unfolds. The most widely-used offline experiments in the GenSLA tradition are acceptability judgement tests (AJT) (Ionin, 2012; Sorace, 1996). Participants rate the acceptability of one (or two) target sentence(s) (usually on a 1-5 Likert scale, where 1 is unacceptable and 5 acceptable), which may be often be preceded by a context which biases for one interpretation (e.g., imperfect tense in (1b) over preterit tense in (1a)) (Domínguez, Arche & Myles, 2017). The simultaneous presentation of two target sentences, each followed by a Likert scale, allows to test a key issue in GenSLA, optionality, as it is often the case that learners' ILGs tolerate both structures to vaying degrees. The Self-Paced Reading Task (SPRT) (Roberts, 2012) is a typical online experiment that measures the milliseconds it takes readers to parse (i.e., syntactically analyse) a sentence as it unfolds in real time. For example, learners are presented on a computer screen with sentences like (2) in a word-by-word fashion (marked by '/'). The learner presses the key to advance to the next word. Crucially, when the learner's parser (mental syntactic processor) reaches the pronoun *he*, an ambiguity is created since it could refer to either the subject *John* or the object *Peter*. The computer accurately measures the extra time it takes to process such ambiguities when compared to non-ambiguous sentences. The combined use of offline vs online methods is important for the latest versions of the IH (Sorace, 2011, 2012), which stipulates that, even though near-native learners may show native-like competence/knowledge in offline tasks, their performance/parsing may be non native-like since integrating syntactic with discursive knowledge uses additional cognitive resources.

(1) When Ana was a child she had a very close friend, Amy, and she like to spend a lot of time at her house after school.

(a) Ana estuvo mucho en casa de Ana al salir del colegio. -2 -1 0 +1 +2

(b) Ana estaba mucho en casa de Ana al salir del colegio. -2 -1 0 +1 +2

'Ana was$_{pret/imperf}$ in Amy's house a lot after getting off school'

(2) John / greeted / Peter / as / he / opened / the / door.

LCR		***GenSLA***
tradition		***tradition***
- control		+control
+ natural		- natural
◄·····►		
	Controlled	
Corpora	**production tasks**	**Experiments**
●*Single-task*	●Sentence	*Offline*
L2 corpora	completion tasks	*experiments:*
●*Multi-task*	●Gap filling tasks	●AJT
L2 corpora		●Interpretation
		tasks
		Online
		experiments:
		●SPRT

Table 1. Different research methods used in LCR vs GenSLA

Experimental methods have been favoured in GenSLA for several reasons. Researchers are often interested in linguistic phenomena that may have low frequency in an L2 corpus (known as *construct underrepresentation*). They often need to know what learners accept/produce as a possible ILG, but crucially what they reject or rule out. Additionally, the production of a form in the corpus may not necessarily reflect the learner's competence and the absence of a form does not entail lack of knowledge. Lack of production may be a sampling issue: the form may

be absent due to the learner's use of avoidance strategies, or there may not be enough instances of the form in the corpus so as to assume competence. Experiments tap competence more directly that corpora do.

The use of L2 corpora in GenSLA has been therefore rather limited to date (but see Lardiere, 2007; Myles, 2005, 2015; Rutherford & Thomas, 2001). A notable exception was Patty's corpus, which was used to test the Feature Reassembly Hypothesis (Lardiere, 2007) on an L1 Chinese-L2 English adult immigrant in the USA after 10 years of residence, but the data of this *ad-hoc* case-study corpus are rather limited (4 recordings and 25 emails in total) (Lardiere, 2008). Such case-study corpora lack extrapolability to the learner population. There is therefore a "need to test hypothesis on larger and better constructed databases" (Mendikoetxea, 2014, p. 12) for several reasons (Lozano & Mendikoetxea, 2013; Myles, 2005, 2007, 2015): Large L2 corpora can provide a wider empirical basis against which to test specific hypotheses than previous small-scale expermiental studies. The process of tagging ILG data (see Chapter 6) also serves as an exploratory process that allows the researcher to uncover new factors that might have gone unnoticed in experimental work, which in turn can help the researcher design a new and better informed experiment whose results can be triangulated against corpus data. For example, Mendikoetxea & Lozano (2018) investigated the acquisition of L2 English Subject-Verb inversion, as in (3) (postverbal subjects shown in bold). The corpus data confirmed a classic hypothesis that has been tested in experiments, namely, the Unaccusative Hypothesis (UH). It basically states that subjects can appear postverbally with a set of instransitive verbs called unaccusatives, such as *exist, occur, appear*), but the corpus data also revealed new insights about the nature of the preverbal element (shown in italics), which could take the form of an expletive as grammatical *there* (3a), but also ungrammatical *it* (3b) and null expletives (3c) or even a loco-temporal PP (3d). The corpus findings were, in turn, implemented in an experiment, which provided newer insights into the acquisition of the preverbal element.

(3) a. *There* exist **about two hundred organizations such as Greenpeace, which have increased the number of its members laterly**.

b. … *it* had occurred **many important events**.

c. ... exist **the science technology and the industrialisation**.

d. *In some places* still exist **popularly supported death penalty**.

L2 corpus data can additionally provide natural and rich discursive information that is essential to fully understand phenomena at the interfaces (e.g., information-structure fators like anaphora resolution at the syntax-discourse interface, as will be seen below). If well-designed, a corpus containing precise learner metadata can provide insights into GenSLA issues like near nativeness, age effects and so on, as will become clear in the following paragraphs.

Different L2 corpus software tools have been used in LCR: concordancers, taggers and annotators (this volume: Chapter 6 on corpus annotators and Chapter 7 on concordancers). Since GenSLA researchers test specific hypotheses, they need fine-grained, linguistically-informed tagsets that incorporate the multiple factors that previous research has shown to be relevant, as in the study of anaphora resolution in L2 acquisition (Lozano, 2009b, 2016) (Figure 1) (cf. the subsection on pronouns and anaphora below for details). In GenSLA broad error tagging and even automatic POS tagging have not been used since they may overlook crucial ILG phenomena (e.g., the information status of the anaphor, the number of potential antecedents of the anaphor, the distance between the anaphor and the antecedent, etc). The UAM Corpus Tool software[1] (O'Donnell, 2009) allows researchers to build complex tagsets for manual annotation and to later perform complex statistical contrasts amongst tags in ways which have not been previously possible in (Gen)SLA. For example, researchers can get statistics for all tags in the tagset in Figure 1, or to statistically compare all/some/each of them amongst subcorpora or amongst individual texts, or to get statistics for a combination of tags (e.g., null

pronouns & topic continuity scenarios) and compare them amongst subcorpora (Lozano, n.d. for a practical illustration).

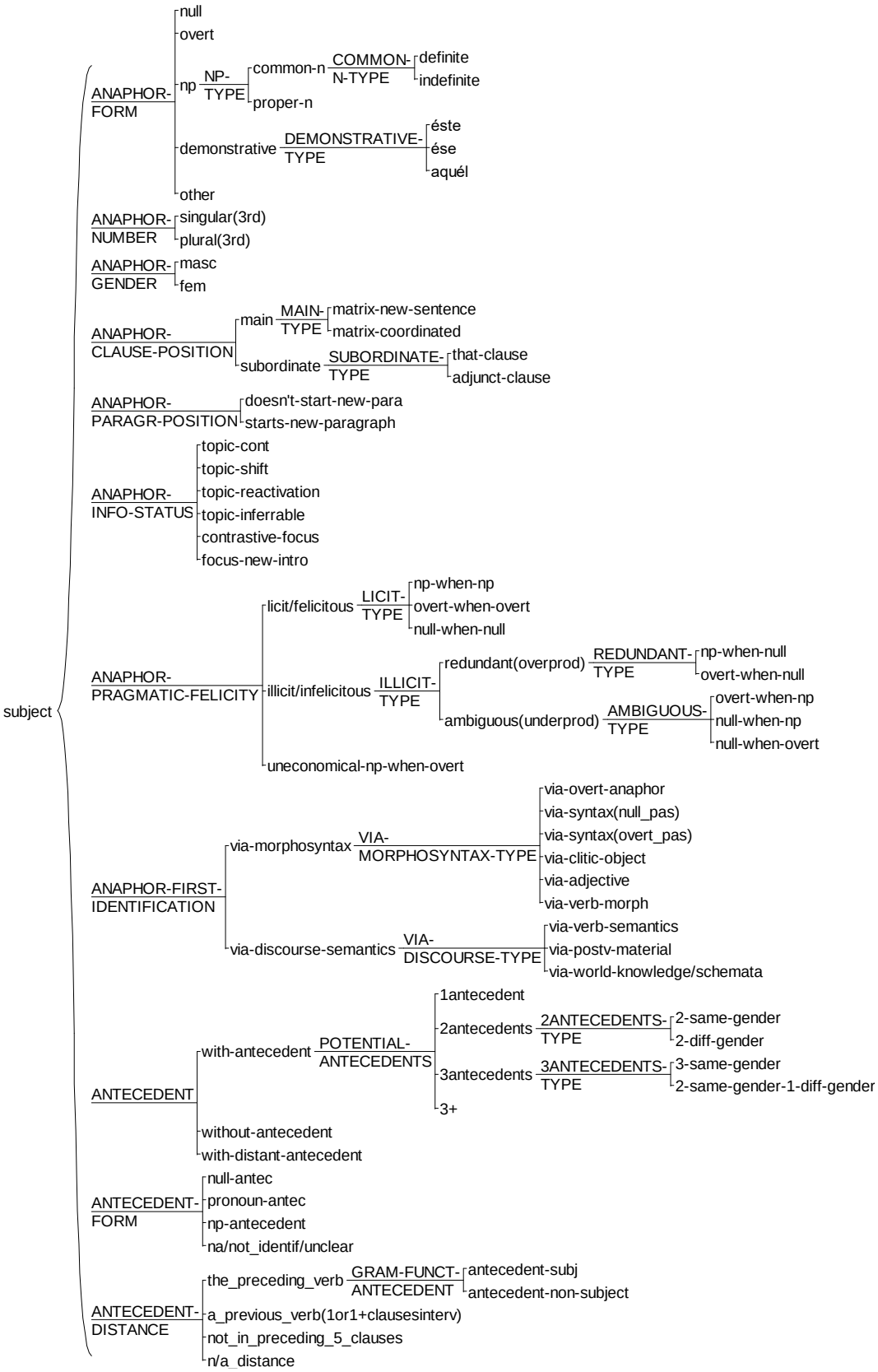

Figure 1. A fine-grained, linguistically-motivated tagset (Lozano 2016)

When designing a corpus (cf. Chapter 5 in this volume for an overview), (Gen)SLA researchers should follow general corpus design principles and collect (Gen)SLA-relevant variables (see Lozano & Mendikoetxea, 2013 for tips), e.g.:

i. Multi-L1 corpora: A variety of learner subcorpora sampling different L1s (ideally with a some L1s differring in terms of parameters/features) will shed light on possible UG influence in case all learners behave similarly irrespective of their L1s.

ii. SLA-motivated learner variables: Recording certain types of learner metadata is crucial to understand some SLA phenomena, e.g.: *age of exposure to the L2* to investigate age-related effects and the Critical Period Hypothesis on learners' deficits with features, as argued above; *length of residence in the L2 target country* check whether immersion and native input may (not) help learners overcome certain residual deficits after prolongued exposure to the L2, as was the case of Patty explained above; the *languages used* by the learner in daily interactions to explore language dominance and attrition effects, as discussed above; and the *proficiency level* of the learners (as measured by a standardised placement test and not by ad-hoc measures like year/grade at school, as often done in LCR) to test developmental hypotheses. These variables help select/filter out learners in theoretically-motivated GenSLA studies of, e.g., ultimate attainment, where it is essential to discriminate near-natives from very advanced learners.

iii. Control corpus: It is essential to use a comparable native control corpus whose design is similar to the L2 corpus (i.e., same tasks, same linguistic variables). The L2 corpus can then be contrasted against the native corpus so as to determine (i) whether native-input frequency (as reflected in the native corpus) shapes the learners' ILGs, since recall that GenSLA has shown that L2 acquisition is not always driven by input frequency alone; (ii) learners' overuse/underuse/misuse of linguistic items when compared to natives, which is central to LCR (Callies, 2015, p. 40) and which may shed light on a

key issue in GenSLA: ultimate attainment in near natives; (iii) whether the observed L2 variability/optionality is either genuine learner variability triggered by the representational/processing deficits explained above or rather a natural reflection of the variability present in the native grammar (i.e., theoretical models may predict a certain phenomenon to be categorical in native grammars but corpus production data may reveal it is variable). Finally, GenSLA researchers may need to use two native corpora as control corpora, i.e., a native corpus of the learners's L1 and a native corpus of the learners' L2 (see section on future directions below for a justification).

In short, corpus data may be complemented with experimental data as they are not mutually exclusive but rather complementary (cf. Chapter 11 in this volume). Triangulating both methods to investigate the same phenomenon represents an advance in (Gen)SLA (cf. the section on future directions below).

Representative corpora and research

A couple of representative GenSLA studies will illustrate how corpus data can inform the acquisition of grammar. See Rankin (2015) and Myles (2015) for general reviews of corpus-based studies on L2 grammar and Lozano and Mendikoetxea (2008, 2010) and Rankin (2009) for specific GenSLA corpus-based studies on word order.

Tense-Aspect Marking

Domínguez et al. (2017) is a GenSLA study that shows how the combined used of corpus data and experimental data provide better insights into aspectual contrasts in L2 acquisition.

Corpus and method: Data were taken from the Spanish Learner Language Oral Corpus (version 2), SPLLOC2[2] http://www.splloc.soton.ac.uk/(Mitchell, Domínguez, Arche, Myles & Marsden, 2008), a spoken corpus of L1 English-L2 Spanish learners at three proficiency levels (beginner/intermediate/advanced according to hours on instruction), plus a comparable Spanish native subcorpus. They compared results from two corpus tasks eliciting different past tense/aspect contrasts (a guided interview about the participant's past, and a more controlled narrative retell task with eliciting prompts) against a contextualised AJT, as in (1) above.

Phenomenon: With past tense, perfective aspect encodes finished events but imperfective aspect encodes unfinished events that can be of three types: continuous (ocurring any number of occasions), habitual (ocurring more than once) and progressive (ocurring once) (Table 2). Spanish assembles this aspectual contrast straightforwardly: perfective features→preterit tense; imperfective features→imperfect tense. But English is not so straightforward: perfective & imperfective (continuous and habitual[3])→past-simple tense; imperfective (progressive)→periphrasis (*be*+verb-*ing*). The task of L1 English-L2 Spanish learners is (i) to acquire that there is an straightforward meaning-form correspondece in Spanish (perfectivity-preterit, imperfectivity-imperfect), (ii) to remap/reassemble the imperfective features (continuous, habitual, progressive), which already exist in their L1 English, onto their new Spanish imperfect verbal morphology. More importantly, learners have to remap the aspectual meanings encoded by the English past simple onto the new Spanish forms (preterit, imperfect).

	Features	English native forms	Spanish native forms	L1 English-L2 Spanish learner forms
PERFECTIVE	[finished] [1 occasion]	Past simple: *Marta **was** ill last sunday*	Preterit: *Marta **estuvo** enferma el domingo*	Preterit: *Marta **estuvo** enferma el domingo*
IMPERFECTIVE	CONTINUOUS [unfinished] [any occasions]	Past simple: *Marta **was** ill (when I visited her)*	Imperfect: *Marta **estaba** enferma (cuando la visité)*	*Preterit / Imperfect: *Marta **estaba**/**estuvo** enferma (cuando la visité)* FINDINGS:[c] •Exper: most problematic •Corpus: problematic
	HABITUAL [unfinished] [>1 occasion]	Past simple: *Marta **sang**[a] in a choir when she was little*	Imperfect: *Marta **cantaba**[a] en un coro cuando era pequeña*	*Preterit / Imperfect: *Marta **cantaba**/**cantó** en un coro cuando era pequeña. FINDINGS:[c] •Exper: least problematic •Corpus: most problematic

	PROGRES SIVE [unfinished] [1 occasion]	***Be*+V-ing**	*Marta* ***was singing*** *when we arrived*		*Marta* ***cantaba***[b] *cuando llegamos*		*Marta* ***cantaba/*cantó*** *cuando llegamos.* FINDINGS:[c] •Exper: problematic •Corpus: least problematic

Table 2. Feature configuration of perfective/imperfective aspect in native English, native Spanish and L1 English-L2 Spanish

Table notes: [a] Periphrastic forms are also used to express habitual imperfective aspect both in English (*Marta would sing/used to sing in a choir*) and in Spanish (*Marta solía cantar en un coro*). [b] Periphrastic forms are also possible in Spanish for progressive imperfective aspect (*Marta estaba cantando cuando llegamos*). [c] For simplicity reasons, findings refer only to the advanced learners.

Predictions: According to the FRH (cf. section on core issues above), learners are expected to have problems when remapping or reassembling the perfective/imperfective aspectual distinction onto their corresponding forms in L2 Spanish since English and Spanish differ in the way this distinction is assembled in morphological forms. Learners will overextend the aspectual meanings associated with English past simple morphology (perfective, imperfective continuous/habitual) to the Spanish preterit morphology.

Results: Imperfect verbal morphology is correctly produced/accepted from early stages, but the array of interpretations (continuous, habitual, progressive) is not completely acquired even at

advanced levels (cf. Table 2, last column). They incorrectly overextend preterit forms (*estuvo, cantó*) to all three imperfective contexts, though the degree of overextension depends on the nature of the task (corpus production vs experimental interpretation task). Overextenstion was highly problematic in continuous contexts in the experiment (i.e., acceptance of ungrammatical *Marta *estuvo enferma cuando la visité*), but in habitual contexts in the corpus (i.e., production of *Marta *cantó en un coro cuando era pequeña*). These mapping problems are predicted by the FRH. Importantly, if only corpus data had been used, the researchers would have concluded that overextension happens only in habitual contexts. If only experimental data had been used, only continuous contexts would have been shown to be highly problematic. In short, the combined used of corpus and experimental data, coupled with theoretically-motivated predictions, provides nuanced insights into aspectual distinctions in SLA.

Pronouns and anaphora

Taking the IH as a starting point (cf. section on core issues above), Lozano (2016) used L2 Spanish corpus data and proposed the PPVH to account for anaphora resolution (AR).

Phenomenon: AR refers to how referential expressions (REs) like overt pronouns (*ella* 'she'), null pronouns (*Ø*) and NPs (*el abogado* 'the lawyer'), shown in bold in (4) and (5), refer to previously mentioned antecedents (in italics) in the discourse. REs can syntactically alternate in subject position (*El abogado/él/Ø tiene problemas*), but such alternation is constrained at the syntax-discourse interface by information structure in native Spanish: in topic continuity the sentential topic is maintained via a null pronoun (*el abogado*$_i$ ... *Ø*$_i$ in (5a) and *Ella*$_i$ … *Ø*$_i$ in (4)), but an overt pronoun represents a topic shift (*el abogado*$_i$ ... *Ella*$_j$ in (5b)), though an NP could also mark a topic shift (*Ella*$_j$ ... *Ø*$_j$ **…** *El abogado*$_i$ in (4)).

(4) *Este abogado*$_i$ tiene sus proprias problemas con *su hija*$_j$. **Ella**$_j$ es una adicta de heroína, y [**Ø**$_j$] le$_j$ llama casi cada día, pidiendo ayuda o dinero. **El abogado**$_i$ casi le$_j$ ha renunciado… [L1 Eng-L2 Spa advanced learner, CEDEL2 corpus]

'This lawyer$_i$ has his own problems with his daughter$_j$. She$_j$ is an addict to heroin, and [she$_j$] calls him$_j$ nearly every day, asking for help or money. The lawyer$_i$ has nearly renounced her$_j$ …

(5) a. El abogado$_i$ tiene problemas con su hija$_j$. [**Ø**$_i$] Buscará una solución.

b. *El abogado*$_i$ tiene problemas con su *hija*$_j$. **Ella**$_j$ buscará una solución.

'The lawyer has problems with his daughter. [He] / She will find a solution.'

Corpus and method: Lozano (2016) used CEDEL2[4] (*Corpus Escrito del Español L2*) (Lozano, 2009a; Lozano & Mendikoetxea, 2013) (N=2,578 total participants). He compared very advanced L1 English-L2 Spanish learners with an equally-designed native Spanish control corpus. Film-retell tasks were annotated with a linguistically informed fine-grained tagset (Figure 1 above) implemented in the UAM Corpus Tool software.

Predictions: The IH predicts that, at the syntax-discourse interface, L2 learners will produce overt (instead of null) pronominal subjects as the default form due to the difficulty of integrating syntactic (licensing of null pronouns) and discursive information (topic continuity vs topic shift). Taking the IH as a starting point, Lozano used natural corpus production data to test the PPVH (cf. section on core issues above).

Results: In topic-continuity contexts, learners differed from natives as they redundantly used overt pronouns, as predicted by the IH, but the corpus data uncovered a more complex picture that previous GenSLA experimental studies had overlooked: the number of potential antecedents in the previous discourse shapes the form of the RE in subtle ways, i.e., (i) in topic

continuity such 'redundant' pronouns are produced with two (but not with only one) competing antecedent; (ii) in topic shift both natives and learners used an overt pronoun more often than an NP when there were two competing antecedents, but with three competing antecedents when an NP is the norm. Learners' showed native-like sensitivity to the number of competing antecedents, but they are more often redundant than ambiguous. The PPVH was proposed to account for these new corpus findings.[5] The GenSLA hypothesis-testing (coupled with a corpus-based hypothesis-finding model) can shed new light on factors that were overlooked in previous experimental studies using GJT and related methodologies.

Future directions

This section discusses key issues and recommendations for future GenSLA/LCR researchers.

The first issue relates to triangulation, which is the standard scientific practice of using different research methods to investigate a phenomenon. Some researchers recently advocate for the triangulation of corpus and experimental data in SLA research (Díaz-Negrillo & Thompson, 2013, p. 21; Gilquin, this volume; Lozano & Mendikoetxea, 2013, p. 89; Tracy-Ventura & Myles, 2015, p. 61), particularly when the phenomenon under investigation is infrequent in the corpus (Callies, 2015, p. 42). Recall that combining corpus and experimental data provides new insights into tense-aspect contrasts (Domínguez et al., 2017; Domínguez, Tracy-Ventura, Arche, Mitchell & Myles, 2013). Mendikoetxea & Lozano (2018) show how corpus and experimental methods can be used to investigate the very same linguistic phenomenon (postverbal subjects) in a cyclic fashion, whereby the new factors uncovered in the corpus (which went undetected in previous experimental studies) can be implemented in a new experiment, whose results may ultimately inform future corpus analyses.

A second issue concerns the use of two native control corpora. Apart from including an equally-designed native control corpus of the learners' target language (as done in CEDEL2 and SPLLOC), a further step is using an additional control corpus. For example, the L1 English-L2 Spanish data from CEDEL2 (version 2.0) can be compared against two control corpora: an L1 Spanish native control subcorpus from CEDEL2 itself and an equally-designed L1 English native control subcorpus from COREFL (Corpus of English as a Foreign Language) (Lozano, Díaz-Negrillo & Callies, 2019). GenSLA researcher can thus (i) use real native production data to determine how features are assembled onto the functional categories in both the learners' native language and in the language they are acquiring; (ii) better understand how the featural makeup of those two languages shapes the featural makeup of the learners' ILGs; (ii) explore other types of L1 influence as well as L2 input influence on the learners' ILG.

Another issue is the design of L1↔L2 bidirectional corpora. This design principle is important since it allows researchers to test the same linguistic phenomenon bidirectionally, e.g., L1 English-L2 Spanish (CEDEL2) vs L1 Spanish-L2 English in WriCLE[6] (Written Corpus of Learner English) (Rollinson & Mendikoetxea, 2010). This allows GenSLA researchers to determine which ILG properties are the result of L1-specific factors vs. those that are universal and therefore observable in both corpora.

An additional design issue is bimodal corpora. Whereas most L2 corpora are written (with some cases of spoken corpora like SPLLOC), a welcome addition to SLA would be the use of equally-designed bimodal corpora (written vs spoken) sampling the very same learners producing both tasks. This allows to test the effect of modality, since it has been claimed that learners' competence is reflected better in spoken than written corpora (Myles, 2015, p. 313), but this is an empirical issue that needs further LCR-based corroboration. Additionally, spoken (dialogic) data are suitable to test interface effects where discourse information (new/old, topic/focus) constrains syntactic choices (syntax-discourse interface).

Another issue relates to the use of developmental L2 corpora. Some of the main aims of GenSLA is to understand (i) how ILGs develop over time (from initial states, through transitional states up to end-states), (ii) how parametric/featural choices develop across stages, (iii) whether native-like competence is eventually attainable (end-states). Longitudinal corpora are ideal to test this, but it is logistically complex to track the development of the same group of learners across the years as their proficiency increases. Cross-sectional L2 corpora are a logistically simpler solution, as in CEDEL2 and SPLLOC. Most traditional L2 English corpora like ICLE sample advanced learners only, but recent attempts like the *Longitudinal Database of Learner English*, LONGDALE represents a truly longitudinal corpus.

An important design issue concerns different types of exposure. There is a need for more L2 corpora sampling (ideally the same) learners under different conditions of exposure, as in LANGSNAP[7] (Languages and Social Networks Abroad Project) (Tracy-Ventura, Huensch & Mitchell, forthcoming), who tracked L1 English-L2 Spanish/French university learners before, during and after a residence abroad programme. For advanced learners, genuine exposure to the L2 during residence abroad is expected to enhance discursive aspects, which may be beneficial for syntax-discourse interface phenomena.

An additional key issue relates to the variety of learners. Many L2 corpora sample homogeneous groups (i.e., university learners studying English degrees and performing similar tasks as part of their curriculum, as in ICLE). This can undermine the balance/representativeness of the corpus (Gilquin, 2015), as findings may not be extrapolable to the entire population of L2 learners. Such corpora may not be sufficient to address theoretical questions like ultimate attainment, which typically require the learner to have resided in the target country for several years to be considered near native. Current GenSLA shows that L2 dominance, which is often the result of the learner's prolongued exposure to (and use of) the L2 in the target country, affects certain peripheral areas (interfaces) of the L1, though these

effects appear not to be irreversible (Chamorro, Sorace & Sturt, 2016; Domínguez, 2013; Sorace, 2004). This is known as L1 attrition but to date there is no publicly available corpus of L1 attriters. Corpora with varied types of bilinguals is ideal, as Gilquin (2015) argues:

> "What is particularly interesting about this corpus [CEDEL2] is that, unlike most learner corpora which are collected in a small number of environments …, speakers of Spanish all over the world were invited to contribute. This results in a wide range of writer profiles, using different varieties of (learner and native) Spanish" (p. 23).

A final issue is the variety of communicative tasks. Argumentative essays have been the mainstream task in many L2 corpora, but different communicative tasks provide learners with opportunities to make relevant linguistic contrasts in a variety of contexts (Callies, 2015; Myles, 2015). Tracy-Ventura & Myles (2015) used a variety of corpus tasks from the SPLLOC corpus and showed that if only a standard descriptive task had been used instead of other highly controlled narrative tasks, certain tense-aspect contrasts would not have shown up in the learner data. Multi-task corpora may therefore provide a better reflection of learners' competence. Unlike previous experimental studies, Lozano (2009b) found that, out of 12 tasks from CEDEL2 (descriptive, narrative and argumentative), learners were shown to exhibit deficits only in those tasks eliciting anaphora resolution contexts in 3rd person singular human contexts, and not in tasks eliciting 1st or 2nd person.

To summarise, hypothesis-testing studies that are backed up by specifically-designed L2 corpus data can provide new insights into key aspects in GenSLA.

Further reading

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 309-332). Cambridge: Cambridge University Press.

Myles, F. (2007). Using electronic corpora in SLA research. In D. Ayoun (Ed.), *Handbook of French Applied Linguistics* (pp. 377-400). Amsterdam: John Benjamins.

Myles (2015) and her earlier version (Myles, 2007) focus on how LCR can best contribute to SLA theory. Myles also discusses what is needed in an L2 corpus for SLA research, which will provide insights for future (Gen)SLA L2 corpus designers.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 65-100). Amsterdam: John Benjamins. Retrieved from https://doi.org/10.1075/scl.59.06loz

Lozano & Mendikoetxea (2013) discuss L2 corpus design principles and variables. They highlight ten basic principles of learner corpus design (taking CEDEL2 as a case in point), as well as a proposal for the type of learner and task variables needed if the corpus intends to answer SLA-relevant questions (cf. also Gilquin's 2015 review of the design of CEDEL2, which also contains additional recommendation on L2 corpus design principles and how to collect the data).

Related topics

Chapter 5. Learner corpus design, collection and transcription

Chatper 6. Annotating learner corpus data

Chapter 7. Analyzing a learner corpus with a concordancer

Chapter 11. Combining learner corpora and experimental methods

Notes

1. http://www.corpustool.com (last accessed 16 December 2019).

2. http://www.splloc.soton.ac.uk (last accessed 16 December 2019).

3. Note that habitual can be also realised periphrastically with *would/used to*.

4. http://cedel2.learnercorpora.com (last accessed 16 December 2019).

5. Recent corpus-informed experimental data confirm the PPVH (Lozano, 2018).

6. http://wricle.learnercorpora.com/ (last accessed 16 December 2019).

7. http://langsnap.soton.ac.uk/ (last accessed 16 December 2019).

References

Callies, M. (2015). Learner Corpus Methodology. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 35-55). Cambridge: Cambridge University Press.

Chamorro, G., Sorace, A., & Sturt, P. (2016). What is the source of L1 attrition? The effect of recent L1 re-exposure on Spanish speakers under L1 attrition. *Bilingualism: Language and Cognition*, *19*(3), 520-532. https://doi.org/10.1017/S1366728915000152

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MASS.: MIT Press.

Chomsky, N. (2000). *The Architecture of Language*. Oxford University Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1-22.

Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language Learning and Development*, *7*, 263-278.

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 9-29). Amsterdam: John Benjamins.

Domínguez, L. (2013). *Understanding Interfaces: Second language acquisition and first language attrition of Spanish subject realization and word order variation* (Vol. 55). Amsterdam: John Benjamins. Retrieved from http://www.jbe-platform.com/content/books/9789027271990

Domínguez, L., Arche, M. J., & Myles, F. (2017). Spanish Imperfect revisited: Exploring L1 influence in the reassembly of imperfective features onto new L2 forms. *Second Language Research*, *33*(4), 431-457. https://doi.org/10.1177/0267658317701991

Domínguez, L., Tracy-Ventura, N., Arche, M. J., Mitchell, R., & Myles, F. (2013). The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition*, *16*(03), 558-577. https://doi.org/10.1017/S1366728912000363

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9-34). Cambridge: Cambridge University Press.

Hawkins, R. (2001). *Second Language Syntax: A Generative Introduction*. Oxford: Blackwell.

Hawkins, R. (2007). The nativist perspective on second language acquisition. *Lingua*, *118*(4), 465-477.

Ionin, T. (2012). Formal Theory-Based Metodologies. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 30-52). Oxford: Wiley-Blackwell.

Lardiere, D. (2007). *Ultimate Attainment in Second Language Acquisition: A Case Study*. Mahwah, N.J: Lawrence Erlbaum Associates.

Lardiere, D. (2008). Feature assembly in second language acquisition. In J. M. Liceras, H. Zobl & H. Goodluck (Eds.), *The Role of Formal Features in Second Language Acquisition* (pp. 106-140). New York: Lawrence Erlbaum Associates.

Lardiere, D. (2009). Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research*, *25*(2), 173-227. https://doi.org/10.1177/0267658308100283

Liceras, J. M., Zobl, H., & Goodluck, H. (Eds.). (2008). *The Role of Formal Features in Second Language Acquisition*. Mahwaw, NJ: Lawrence Erlbaum Associates.

Lozano, C. (2003). *Universal Grammar and focus constraints: The acquisition of pronouns and word order in non-native Spanish* (PhD). University of Essex.

Lozano, C. (2009a). CEDEL2: Corpus Escrito del Español como L2. In C. M. Bretones et al. (Eds.), *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente* (pp. 197-212). Almería: Universidad de Almería.

Lozano, C. (2009b). Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In Y.-I. Leung, N. Snape & M. Sharwood-Smith (Eds.), *Representational Deficits in Second Language Acquisition* (pp. 127-166). Amsterdam: John Benjamins. https://doi.org/10.1075/lald.47.09loz

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus. In M. Alonso Ramos (Ed.), *Spanish Learner Corpus Research: Current Trends and Future Perspectives* (pp. 235-265). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.78.09loz

Lozano, C. (2018). The development of anaphora resolution at the syntax-discourse interface: pronominal subjects in Greek learners of Spanish. *Journal of Psycholinguistic Research*, *47*(2), 411-430. https://doi.org/10.1007/s10936-017-9541-8

Lozano, C. (n.d.). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. In M. Cruz Piñol (Ed.), *E-Research y español LE/L2: Investigar en la era de las tecnologías*. New York: Routledge.

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2019). Designing and compiling a learner corpus of written and spoken narratives: COREFL. In C. Bongartz & J. Torregrossa (Eds.), *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*. Frankfurt: Peter Lang.

Lozano, C., & Mendikoetxea, A. (2008). Postverbal subjects at the interfaces in Spanish and Italian learners of L2 English: a corpus analysis. In G. Gilquin, S. Papp & M. B. Díez-

Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 85-125). Amsterdam: Brill-Rodopi. https://doi.org/10.1163/9789401206204_005

Lozano, C., & Mendikoetxea, A. (2010). Interface conditions on postverbal subjects: A corpus study of L2 English. *Bilingualism: Language and Cognition*, *13*(4), 475-497. http://dx.doi.org/10.1017/S1366728909990538

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 65-100). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59.06loz

Mendikoetxea, A. (2014). Corpus-based research in second language Spanish. In K. L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp. 11-29). Oxford: Wiley-Blackwell.

Mendikoetxea, A., & Lozano, C. (2018). From corpora to experiments: methodological triangulation in the study of word order at the interfaces in adult late bilinguals (L2 learners). *Journal of Psycholinguistic Research*, *47*(4), 871-898. https://doi.org/10.1007/s10936-018-9560-0

Mitchell, R., Domínguez, L., Arche, M., Myles, F., & Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. In L. Roberts, F. Myles & A. David (Eds.), *EUROSLA Yearbook 8* (pp. 287-304). Amsterdam: John Benjamins.

Montalbetti, M. (1986). How pro is it? In *Studies in Romance Linguistics* (pp. 137-152). Dordrecht: Foris.

Montrul, S. (2011). Multiple interfaces and incomplete acquisition. *Lingua*, *121*(4), 591-604. https://doi.org/10.1016/j.lingua.2010.05.006

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), 373-391.

Myles, F. (2007). Using electronic corpora in SLA research. In D. Ayoun (Ed.), *Handbook of French Applied Linguistics* (pp. 377-400). Amsterdam: John Benjamins.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 309-332). Cambridge: Cambridge University Press.

O'Donnell, M. (2009). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M. Bretones et al. (Eds.), *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente* (pp. 1433-1447). Almería: Universidad de Almería.

Rankin, T. (2009). Verb second in advanced L2 English: A learner corpus study. In M. Bowles, T. Ionin, S. Montrul & A. Tremblay (Eds.), *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference*. Somerville, MA: Cascadilla Proceedings Project.

Rankin, T. (2015). Learner corpora and grammar. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 231-254). Cambridge: Cambridge University Press.

Roberts, L. (2012). Psycholinguistic techniques and resources in second language acquisition research. *Second Language Research*, *28*(1), 113-127. https://doi.org/10.1177/0267658311418416

Rollinson, P., & Mendikoetxea, A. (2010). Learner corpora and second language acquisition: Introducing WRICLE. In J. L. Bueno Alonso, D. Gonzáliz Álvarez, U. Kirsten Torrado, A. E. Martínez Insua, J. Pérez-Guerra, E. Rama Martínez & R. Rodríguez Vazquez (Eds.), *Analizar datos > Describir variación / Analysing Data > Describing Variation* (pp. 1-12). Vigo: Universidade de Vigo (Servizo de Publicacións).

Rothman, J., & Slabakova, R. (2011). The mind-context divide: On acquisition at the linguistic interfaces. *Lingua, 121*, 568-576.

Rothman, J., & Slabakova, R. (2018). The generative approach to SLA and its place in modern second language studies. *Studies in Second Language Acquisition*, *40*(2), 417-442. https://doi.org/10.1017/S0272263117000134

Rutherford, W., & Thomas, M. (2001). The Child Language Data Exchange System in research on second language acquisition. *Second Language Research*, *17*(2), 195-212.

Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W. C. Ritchtie & T. K. Bhatia (Eds.), *Handbook of Second Language Acquisition*. London: Academic Press.

Sorace, A. (2000). Syntactic optionality in non-native grammars. *Second Language Research*, *16*(2), 93-102.

Sorace, A. (2004). Native language attrition and developmental instability at the syntax-discourse interface: Data, interpretations and methods. *Bilingualism: Language and Cognition*, *7*(2), 143-145.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism, 1*(1), 1-33.

Sorace, A. (2012). Pinning down the concept of "interface" in bilingualism: a reply to peer commentaries. *Linguistic Approaches to Bilingualism*, *2*(2), 209-216.

Tracy-Ventura, N., Huensch, A., & Mitchell, R. (forthcoming). Understanding the evolution of L2 lexical diversity over time: The contribution of a longitudinal learner corpus. In B. Le Bruyn & M. Paquot (Eds.), *Second Language Acquisition and Learner Corpora*. Cambridge: Cambridge University Press.

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, *1*(1), 58-95. https://doi.org/10.1075/ijlcr.1.1.03tra

White, L. (1989). *Universal Grammar and Second Language Acquisition*. Amsterdam: John Benjamins.

White, L. (2003). *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

White, L. (2011). Second language acquisition at the interfaces. *Lingua*, *121*(4), 577-590. https://doi.org/16/j.lingua.2010.05.005

White, L. (2012). Universal Grammar, crosslinguistic variation and second language acquisition. *Language Teaching*, *45*(3), 309-328. https://doi.org/10.1017/S0261444812000146