COVER PAGE

Running head: Comparing Learner Corpora

Authors:

Sandra C. Deshors, Michigan State University (ORCID ID: 0000-0003-0726-2102)

Stefan Th. Gries, University of California, Santa Barbara & Justus Liebig University Giessen (ORCID ID: 0000-0002-6497-3958)

Short bios:

Sandra C. Deshors

Sandra C. Deshors is Assistant Professor in the Second Language Studies Ph.D. program at Michigan State University. Her research, which specializes in quantitative corpus-based approaches to learner language, contrasts English as a Foreign Language (EFL), English as a Second Language (ESL) and World Englishes at large. Theoretically, her work is anchored in the usage-based theoretical framework and recognizes a correlation between speakers' mental knowledge of linguistic items and their uses in grammatical contexts.

Stefan Th. Gries

Stefan Th. Gries is a Professor in the Department of Linguistics of the University of California, Santa Barbara and Chair of English Linguistics (Corpus Linguistics with a focus on quantitative methods) in the Department of English, Justus-Liebig-Universität Giessen.

He is a quantitative corpus linguist with an interest in cognitive/usage-based as well as psycholinguistic applications.

Comparing Learner Corpora

Sandra C. Deshors
Michigan State University

Stefan Th. Gries
University of California, Santa Barbara
& Justus Liebig University Giessen

Abstract

This chapter provides a survey of past, present, and future directions of corpus-based comparisons of learner data in the fields of Learner Corpus Research (LCR) and Second Language Acquisition (SLA). First, we discuss the notion of comparison theoretically by situating existing work in LCR (i) within frameworks such as Contrastive Interlanguage Analysis and the Integrated Contrastive Model and (ii) in relation to SLA research. Then, we discuss the notion methodologically by presenting main research methods that facilitate the comparison of learner data. Finally, we offer guidelines on how to apply current theoretical and methodological frameworks and we make recommendations for more complete annotation schemes, better control and the development of comprehensive statistical analyses.

Keywords: Comparability of learner corpora, Integrated Contrastive Model (ICM), Contrastive Interlanguage Analysis (CIA), contrastive corpus tools, quantitative/statistical techniques

1 Introduction

The comparison of learner data is a fundamental notion in the fields of Learner Corpus Research (LCR) and Second Language Acquisition (SLA). Indeed, comparing learner data is important for two main reasons. First, by comparing how the use of a given second/foreign language (e.g. English) by a particular learner population (e.g. French learners of English) differs from how native speakers use that language, researchers can explore how the learner variety (or *interlanguage*) differs from the native variety and to what extent observed differences in the learners' output can tell us something about learners' systematic knowledge of their interlanguage. Second, by comparing how different learner populations (e.g. learners with different native language backgrounds) use a common second language, researchers can explore to what extent learners' native language influences their respective interlanguage. Put differently, researchers can capture and understand the forces that drive cross-linguistic transfer during second language production.[1] Importantly, however, although the notion of comparing learner data is central to both LCR and SLA, the two fields have approached the notion differently. In LCR, comparability has been at the core of the methodological framework(s) upon which the entire research field has developed over the past twenty years. Further, the notion has also been at the forefront of the design and compilation of major learner corpora and the development of gradually more and more sophisticated statistical approaches. Already back in the late 1990s when the first large-scale learner corpus was developed, the *International Corpus of Learner English* (ICLE; ICLEv1: Granger et al. 2002; ICLEv2: 2009), comparability across learner Englishes was a central part of the corpus design as scholars recognized that "the main innovative aspect of ICLE is the systematic approach to corpus design and the compilation of **comparable** sub-corpora produced by learners with a wide range of mother-tongue backgrounds" (Hasselgård and Johansson 2011: 37;

our emphasis). However, maximizing the comparability of learner corpora is a complex task that calls for scholars' attention at all stages of corpus research. In this context, and to increase the reliability of learner corpora comparisons, theoretical frameworks such as the Contrastive Interlanguage Analysis (CIA; i.e. comparisons of native vs. non-native language and/or comparisons of non-native varieties) and the Integrated Contrastive Model (ICM; i.e. combination of a contrastive analysis that compares original data in different native languages and CIA), which we discuss below, were developed and widely applied within the learner corpus research community mainly for the description of learner language. However, amongst SLA scholars, those two frameworks have not been at the center of attention (see for example, Ellis et al. 2016 and Ellis & Ferreira-Junior 2009, scholars who do not represent 'mainstream' LCR in the sense that although they still use corpus comparison as a main research method, they have done so without necessarily referring to CIA/ICM). Further, in the SLA community, the two frameworks have met some resistance, particularly with regard to comparability and the notion of normative standard. For instance, Hunston (2002) and Larsen-Freeman (2014) have questioned the validity of comparing interlanguage varieties (ILs) with a target language (TL; i.e. a native norm) on the grounds that such comparisons suffer from a 'comparative fallacy': Comparisons with the TL can seriously hinder the description of the IL (Bley-Vroman 1983: 2) because ILs have been argued to be linguistic systems that should be described "in their own terms" (Selinker 2014: 230). However, as noted in Paquot (2007), a large proportion of SLA research has nonetheless succumbed to the 'comparative fallacy' (see Lakshmanan & Selinker 2001 and Firth & Wagner 1997). In addition, Larsen-Freeman (2014) objects to how TL vs. IL comparisons imply that learners are deficient speakers; and Hunston (2002) criticizes LCR for assuming a native-speaker norm that learners would target with their IL, which does not quite align with Selinker's notion of describing IL systems in their own terms. That being said, much of SLA research compares learners of different proficiency levels and is small scale. Compared to existing

LCR work though, fewer SLA studies include learners of different native language backgrounds learning the same target language (Granger 2009; however, see McManus 2015 for an example of an SLA study that does distinguish between L1 groups).

In this context, this chapter explores theoretical and methodological issues related to comparing learner data with a view to, first, highlight how corpora allow us to analyze larger data sets and, second, how, in this regard, SLA can learn from LCR.[2] We begin the chapter by presenting the Integrated Contrastive Model (ICM) and the Contrastive Interlanguage Analysis (CIA) methodological frameworks upon which the field of Learner Corpus Research has developed (Section 2). Then, we present main research methods and tools scholars have used within the ICM and CIA traditions (Section 3). We continue by discussing representative research approaches trends in LCR including the description and clustering of learner language varieties as well as the prediction of learners' linguistic choices (Section 4). Finally, we end with guidelines for future directions on how to ensure greater, more reliable comparisons of learner corpora.

2 Core issues and topics

2.1 Theoretical frameworks at the core of Learner Corpus Research: the Integrated Contrastive Model (ICM) and Contrastive Interlanguage Analysis (CIA)

For the past two decades, LCR has developed around two main and related methodological frameworks that have shaped the field by establishing principled approaches to comparing learner corpora, Granger's (1996) ICM and CIA approaches. As for the CIA, Hasselgård and Johansson (2011: 57) argue that the approach "has turned out to be a fruitful paradigm"; and as for the ICM, Gilquin (2000/2001: 123) notes that the framework "has undoubtedly much to offer to anyone

interested in SLA". The success of both the CIA and ICM lies in their ability to help scholars characterize individual interlanguage (IL) varieties through the use of automatic and semi-automatic computerized tools. More specifically, with CIA and ICM, the field has witnessed the emergence of studies highlighting general lexical or morpho-syntactic behavioral tendencies within interlanguage varieties. Based on those tendencies, scholars have been able to identify (dis)similarities between different learner populations (see Hasselgård and Johansson 2011 for a recent review of the field).

In essence, together, the frameworks capture linguistic patterns that allow researchers to better distinguish the linguistic systems of learner language from those of native language as well as those of different learner language varieties. The frameworks are related in that CIA is a part of the ICM framework, as illustrated in Figure 1, but they are also different. First, they serve different purposes: while the ICM mainly captures cases of cross-linguistic transfer, CIA serves to explore (individual) learner varieties. Second and consequently, they differ with regard to the type of (learner) language comparisons they involve.

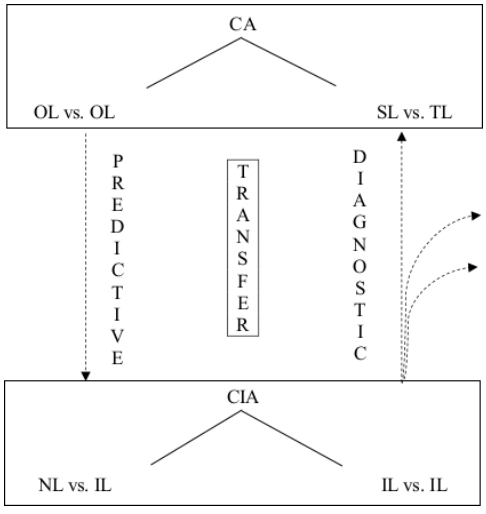

Figure 1 Integrated Contrastive Model (borrowed from Gilquin 2000/2001)[3]

As shown in Figure 1, the ICM combines Contrastive Analysis (CA; i.e. the comparison of source and translated data on the basis of translation corpora) in the upper part of the figure and CIA in the lower part of the figure. In Granger's (1996: 46) words, "[t]he [ICM] model involves constant to-ing and fro-ing between CA and CIA. CA data helps analysts to formulate predictions about interlanguage which can be checked against CIA data […] Conversely, CIA results can only be reliably interpreted as being evidence of transfer if supported by clear CA descriptions". As such, the ICM framework targets the notion of transfer as similarities between the learner's behavior in interlanguage and his/her native language help scholars identify cases of transfer (Gilquin 2008). The CIA, by contrast, involves two major types of comparisons: (i) NL vs. IL, i.e. comparison of native language and interlanguage and (ii) IL vs. IL, i.e. comparison of different interlanguages (Granger 1998: 12).

With regard to (i), Granger (2015: 5) explains that, with this more popular branch of CIA, comparisons of a TL with IL can help reveal overuse that may indicate misuses: "For example, the overuse of *on the contrary* by French learners of English results from a faulty one-to-one equivalence with the French connector *au contraire*" (Granger 2015: 5; more on over- and underuses below).

With regard to (ii), IL vs. IL comparisons, Granger (1993: 60) argues

> [i]n order to be able to distinguish those features of L2 English [or any other natural language] that were L1- dependent, i.e. the result of transfer from the mother tongue, from those which were common to all learners, irrespective of mother tongue, i.e. the cross-linguistic invariants, it [is] essential to enlarge the corpus and include learners from different language backgrounds (Granger 1993: 60).

CIA introduced a type of comparison that served as "a particularly apt basis for a **quantificational** contrastive typology of a number of English ILs [or any other natural language]" (Granger 1998: 12; our emphasis), something that since then has gradually become characteristic (if not a trademark) of LCR: Both the ICM and CIA frameworks assume that distributional patterns of formal elements in IL help us describe and distinguish individual types of IL and better understand why learners shape it the way they do. This view is based on the general assumption within the LCR community that, much like general usage-based theory in linguistics (see Chap. 14), L2 learning is probabilistic in nature: Acquiring a second/foreign language involves increasing knowledge of the frequency of (co-)occurrence of linguistic items in the TL, and distributional differences of formal elements in native and learner language allow researchers to capture traces of non-nativeness (Granger 2004), which is why over-/underuses of linguistic items in IL are central in the field. More specifically, for Granger (2002: 132), identifying over-/underuses helps "bring out the words, phrases, grammatical items or syntactic structures that are either over- or under-used by learners", which is relevant because over- or underuses of formal elements in interlanguage contribute to the "foreign-soundingness […] even in the absence of downright errors" (Granger 2004: 132). Despite their prominence in LCR, criticisms of these notions will be brought up later.

Undeniably, the over-/underuse methodological approach has resulted in numerous descriptive accounts of ILs as well as criticism from some SLA researchers (for instance, the over-/underuses of linguistic items in a learner corpus may not be enough to explain how learner language changes over time, which is a central issue in SLA). However, it has arguably contributed to the development of LCR as a scientific discipline. At the same time, a growing number of LCR scholars have begun to gradually move away from a mere form-based study of over-/underuse towards more context-sensitive (broadly defined, see below) ways of studying quantitatively the complexity of NL vs. IL comparisons (especially when coming from a usage-based theoretical perspective). This analytical

shift has led scholars to adopt more sophisticated statistical approaches such as cluster analysis, correspondence analysis, and logistic regression modeling (see Section 3). This recent development reflects an important effort within the LCR community to harness – rather than move away from – the full potential of the ICM and CIA frameworks and their benefits as theoretical concepts by adopting state-of-the-art methodological approaches and statistical techniques. Indeed, ICM and CIA can include quantitative techniques that not only allow scholars to compare learner corpora descriptively but predictively as well and in an explanatory fashion, which helps in response to criticism from SLA researchers (see Chap. 10).

2.2 Comparison configurations in CIA: some limitations and a second-generation framework

The CIA framework does not require that scholars use a specific native variety as the norm and, fittingly, the notion that any *single* English variety can serve as *the one* native norm has become questionable. This development was recently shown by as Gilquin (2018) in a study of whether American English is a more important source of influence than British English for the other varieties of English (including English as an institutionalized second language and English as a foreign language). Overall, the results point towards a global influence of American English, but also show that varieties are not necessarily homogeneous in this respect and that more local contextual factors may affect the degree of American and/or British influence.

In an attempt to address the above points of criticism, the original CIA framework (CIA^1) was recently revised (CIA^2) by introducing a larger number of reference points against which learner data can be set/compared and broadening its scope to include not just English as a Foreign Language (EFL) varieties, but also English as a Second Language (ESL) varieties and English as a Lingua Franca (see Granger (2015) for a detailed description of CIA^2 and see Crosthwaite et al. (2016) for an example of a study based on the revised CIA framework)[4]. With regard to over- and underuses,

CIA[1] underwent a change in terminology as over- and underuses have become over- and underrepresentations (Granger 2015). Despite these changes, though, to a large extent, the framework remains similar.

2.3 Applying the ICM and CIA frameworks: corpus comparison, comparisons across learner varieties and comparisons across language-production modes

In this section, we consider the application of ICM and CIA in various contexts of L2 uses. Generally, applying these methodological frameworks is not as straightforward as one may think. In order to ensure comparability of native and learner language, various corpus-external and speaker-related factors need to be considered; these include (i) the comparability of corpora in terms of their architecture and their types of data, (ii) the comparability of language production characteristics such as modes (i.e. speech vs. writing) but also contextual characteristics such as genre, formality and purpose of the communication and topic, and (iii) interspeaker variation triggered by factors such as gender, age, regional affiliation, socioeconomic background, cultural background, as well as factors involving aptitude, motivation, proficiency, and others (see, e.g., Gablasova, Brezina and McEnery 2017). Within the ICM and CIA, it would be crucial that contrasted corpora are similar in enough of the above aspects so that observed differences across native speakers and learners can confidently be attributed to the language varieties investigated.

With regard to (i), comparability of corpora, the design and architecture of the ICLE provides a relatively good example of a corpus set up for sound comparisons of IL varieties. [5] As a corpus of IL varieties, ICLE (v2) includes 16 IL varieties. To enhance comparability, all ICLE data were collected all over the world according to the same criteria related to age (young adults of approximately 20 years of age), learning context (studying English in a non-English speaking environment), proficiency level (*advanced* as defined by their seniority in an undergraduate degree

in English), medium of communication (writing) and (argumentative/literary) essay writing. In addition, variables such as sex of the participant, mother tongue, region, other foreign languages, practical experience (i.e. number of years of English teaching), topic of discussion and task setting (Granger 1998) were included, too. While those features, which are part of ICLE's metadata, are also accounted for in corpora of spoken EFL such as the Louvain Corpus of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010), making EFL comparisons across IL varieties and language-production modes relatively straightforward, comparing EFL and ESL quickly becomes complex when one needs to contrast corpora whose respective architecture and design vary widely; comparisons become especially tricky given that corpora often differ considerably in the amount, resolution, and precision of their metadata (see Chap. 6). Further, more recent learner corpora have adopted different designs and have strived to include more characteristics in their annotation; see the relevant chapter(s) on corpus design and annotation in this volume.

To date, the only corpus of learner English (i.e. EFL) that allows for full comparison with ESL varieties is the Corpus of Dutch English (Edwards 2016), modeled on the design of the International Corpus of English (ICE; Greenbaum and Nelson 1996), a corpus collected to contrast varieties of English worldwide. This, in turn, raises the important issue of text-type comparability, which explains why studies such as Deshors (2016), which contrast the two types of learners (i.e. EFL vs. ESL learners), are limited to using a small portion of the ICE data, namely the student writing sub-part most comparable to ICLE's argumentative texts. Given its importance, this aspect of learner corpora comparisons was integrated into the revised version of the CIA framework in terms of *diatypic* variables (Granger 2015). While the operationalization of such variables remains to be made clear, in essence, they are recognized to be "essential to ensure text-type comparability" (Granger 2015: 17).

Moving on to (ii), language production characteristics (e.g. language production modes (speech vs. writing), and contextual characteristics), which are essential to sound compari- sons of learner corpora. As Gablasova, Brezina and McEnery (2017: 137) note, "[w]hen deciding whether two corpora can be meaningfully compared, the likelihood of occurrence of the target linguistic features also has to be considered with respect to the nature of the linguistic data in the corpus". With regard to language production modes specifically, Biber's (1988, *passim*) multidi- mensional analysis has shown that the two language modes attract different lexico-grammatical features (see also McCarthy and Carter 2001). In the case of learner language, however, this distinction between language modes is not always as clear-cut as it is in native language: Gilquin and Paquot (2008) have shown that EFL learners experience difficulties distinguishing between the two modes. In writing, for instance, their uses of linguistic features tend to be more typical of speech than of academic prose, suggesting that they are largely unaware of, or unfamiliar with, register differences. Further, Deshors and Gries (2015: 132) note "without a mode distinction, one cannot be sure that observed pattern differences across corpora are due to variation across varieties rather than registers", which means that comparing language production modes in L2 requires ensuring that, linguistically, the features selected for analysis can be compared meaningfully across modes and that remaining potential differences of modes and genres are accounted for statistically (see Deshors and Gries 2015).

With regard to genre variation, this is not an aspect that can be said to have *really* been explored within the bounds of LCR or SLA, mainly due to the fact that most learner corpora tend to consist of argumentative essays (in written corpora) and academic (in the sense of 'being conducted in university contexts') interviews (in spoken data) as opposed to representing a variety of different genres. Therefore, existing learner corpora are much less representative of the wide range of written and spoken genres found in large-scale corpora of native English such as the ICE. This means that

it is hard to assess whether or to what extent linguistic patterns in IL are (also) influenced by genre, as pointed out by Paquot and Biber (2015), and the lack of alignment regarding genres, tasks, etc. in the design of EFL, ESL and English as a Native Language (ENL) corpora has led scholars to restrict their data to the 'student writing' or 'academic writing' subsections of large-scale corpora such as ICE when contrasting EFL with ESL and ENL data.

As for the final point, (iii) inter-speaker variation, this is an extremely important, yet largely understudied, aspect of LCR and SLA. In a nutshell, accounting for inter-speaker variation means including or controlling for speaker-specific characteristics that may affect speakers' uses of language such as personality, language aptitude, motivation, proficiency, and others (see Gries 2018 for a programmatic discussion). The importance of examining inter-learner variation lies in that it provides a way to "determine to what extent the quantitative summary (e.g., a measure of central tendency such as the mean) can be used to represent the language produced by individual users in the corpus" (Gablasova, Brezina and McEnery 2017: 138). Take for example, the case of learners' proficiency levels, an aspect of learner corpus comparisons that is currently gaining much attention because of how it provides an important statistical control variable. As such, it is often used as a main predictor and/or a dependent variable. Although the notion of learner proficiency has never been completely absent from conversations on comparing learner corpora (e.g. the ICLE was designed with a focus on advanced proficiency levels in mind), it has not always been rigorously and comparably operationalized. For instance, Callies, Díez-Bedmar and Zaytseva (2014) note that the ways in which 'advancedness' has been operationalized in published research differ considerably and, even now, the effects of varying levels of proficiency among subjects remain relatively unknown, thereby to some degree weakening learner corpus comparisons. Thus, developing and operationalizing corpus-based indicators of non-native language proficiency has become a fast-growing branch of LCR (Gablasova, Brezina and McEnery 2017; also see Callies and Götz (2015),

Callies, Díez-Bedmar and Zaytseva (2014), and, in the context of regression-based methodological approaches, see Gries and Deshors 2015 and Chap. 10). Finally, it should be noted that the development of corpora such as the MERLIN corpus (https://www.merlin-platform.eu/) and the Trinity Lancaster Corpus (Gablasova et al. 2015), which include metadata on speakers' proficiency levels based on the Common European Framework of Reference for Languages, could allow scholars to account for speakers' proficiency systematically and in a fine-grained fashion (see McEnery et al. (2019) for a special issue of the *International Journal of Learner Corpus Research* on 'Corpus-based Approaches to Spoken L2 Production: Evidence from the Trinity Lancaster Corpus').

3 Main research methods

Methodologically, learner language research both in LCR and SLA has involved the development and application of various approaches that we discuss below including frequency counts, association measures, hierarchical cluster analysis and regression-based statistical modeling techniques. However, in LCR, perhaps more than in SLA, methodological developments tend to be closely related to how learner corpus linguists have recently shifted their analytical approach from form-based to context-based analyses to explore how and to what extent usage-based theoretical frameworks can help us better understand learner language as a linguistic system (see recent research conducted by Gries and colleagues as well as N. Ellis and colleagues and which has had significant methodological implications for LCR). Since the early days of LCR, three main approaches have emerged in the type of linguistic patterns that scholars have investigated: (i) early form-based approaches to learner language focused on isolated linguistic items and their over- or underuse in IL compared to a TL; (ii) influence from research in dialectology led to typological approaches towards

comparing learner corpora involving comparisons of IL and SL varieties based on catalogs of linguistic items (e.g. Szmrecsanyi and Kortmann 2011); and (iii) multifactorial methodologies anchored in constructionist/usage-based theoretical frameworks led to studies contrasting IL and native varieties by focusing on the co-occurrence patterns of linguistic features within grammatical constructions and across learner corpora. Over time, those analytical approaches to learner language have required of analysts to use increasingly sophisticated methods for both the extraction of complex linguistic patterns and the subsequent statistical analysis of richly annotated data.

Focusing on the former, corpus tools such as user-friendly concordance software (e.g. Wordsmith Tools (Scott 2017), AntConc (Anthony 2010)) have played an important role in LCR by allowing fast data-extraction and computation of relatively simple statistics such as frequency counts and type-token ratios (see Chap. 7). Granger (2015: 5) even argued that "[p]atterns of over- and underuse of linguistic features can readily be identified with the appropriate software tools and methods and provide impetus for further analysis"; see also Gilquin (2015) for discussion of over- and underuse and observed (dis)similarities between institutionalized second-language varieties of English and foreign varieties of English in the areas of syntax, lexis, phraseology and pragmatics. However, as the linguistic phenomena being studied became more complex, the limitations of such approaches became apparent and began to require other, more sophisticated approaches. Therefore, while some research goals are attainable with current corpus tools in terms of extraction (e.g. regular expressions), many more complex studies now involve fine-grained annotation schemes and programming languages such as R or Python to extract (and then statistically analyze) complex linguistic patterns. Such fine-grained annotation schemes have important repercussions on how (complex) comparisons of learner varieties are computed.

Focusing on the statistical side of things, LCR, just like other corpus research, relies on the distributional hypothesis, i.e. the notion that the distribution of words and constructions on their own

and with other linguistic elements reveals something about their functions and/or processing characteristics as well as, potentially, the minds of the speakers whose language production is studied – in LCR, obviously L2 learners' knowledge of a target language: "The frequency with which speakers use linguistic features can provide us with an insight into the state of their interlanguage and is a first step in the study of what motivates the use (or avoidance) of these features in their language" (Gablasova, Brezina, and McEnery 2017: 135).

One can identify different ways of approaching distributional information for the ultimate purpose of comparing learner corpora: for instance, amongst others (i) frequency counts for automatic profiling of learner varieties (e.g. Granger and Rayson 1998); (ii) association measures to pinpoint linguistic features most strongly associated with grammatical constructions in L2 (e.g. Martinez-Garcia and Wulff 2012); (iii) hierarchical cluster analysis to group together non-native varieties that are typologically similar (Szmrecsanyi and Kortmann 2011); (iv) regression-based statistical models to predict learner language (such as MuPDAR; Gries and Deshors 2015).[6] Crucially, although these different statistical approaches all fit within the methodological frameworks of ICM and CIA, they do not abide by the over-/underuse characteristic of the frameworks: ICM/CIA have undergone a shift in the type of quantitative techniques that are used to contrast learner varieties. While ICM/CIA were designed with largely descriptive objectives, today they provide the background for much more advanced quantitative analysis of learners' linguistic choices that allow for testing specific hypotheses about learner language (see Gries and colleagues, *passim*; see Chap. 10, for in-depth illustration and discussion on statistical analyses of learner corpora). This development should not be underestimated as, paradoxically, it speaks to both the potential and the limitation of the ICM and CIA frameworks today. Finally, the above notwithstanding, it is important to keep in mind that even though quantitative approaches have dominated LCR overall, qualitative corpus-based comparisons of learner data should not be

underestimated: regarding the key notions of over- and underuse, for instance, the relevant literature stresses that frequency alone is not sufficient, but that usage needs to be considered as well (e.g. Rosen 2018).

4 Representative corpora and research

Much representative research comparing corpora is based on corpora such as the International Corpus of Learner English (ICLE), the Louvain Corpus of Native English Conversations (LOCNEC), Louvain Corpus of Native English Essays (LOCNESS) and Louvain International Database of Spoken English Interlanguage (LINDSEI), all compiled at the Center for English Linguistics at the Université Catholique de Louvain (UCL) in collaboration with a network of research centers worldwide. Generally, all corpora present data produced by university students at an upper-intermediate to advanced proficiency level in English as a foreign language or by students approaching university entrance, in the case of the native data.

Using these and other corpora, several different research trends in learner corpus research have emerged, including (i) describing and profiling interlanguage varieties, (ii) clustering learner varieties and (iii) predicting learners' linguistic choices. With regard to (i), Granger and Rayson (1998) is an example of how frequency counts can help CIA researchers identify salient lexical behavior in L2 through automated profiling to "form a quick picture" of individual learner populations (Granger and Rayson 1998: 131) and then compare the lexical profiles of those populations. This can, for instance, help identify traces of cross-linguistic transfer. Contrasting French-English IL and native English using ICLE and LOCNESS, Granger and Rayson (1998) first selected a number of word categories (e.g. nouns, adjectives, prepositions, conjunctions) and then

computed frequency counts within each word category to draw a usage profile for individual word categories within a particular learner variety and for each investigated word category, which helps identify linguistic usage patterns characteristic of (individual) learner varieties; in addition, one can assess which members of that category learners over-/underuse compared to a native norm. Despite its statistical simplicity, this type of contrastive approach accounts for the permeable nature of interlanguage systems while offering a coarse-grained, holistic picture of how learner corpora compare.

With regard to (ii), clustering learner varieties, this approach helps explore degrees of cross-varietal (dis)similarities in the uses of particular linguistic items by speakers of different language populations and based on a (large) number of contextual clues. More advanced than Granger and Rayson's (1998) automated profiling technique, this approach involves computing behavioral profiles (comprehensive inventories of elements that co-occur with a word within the confines of a single clause or sentence in actual speech or writing (see Divjak and Gries 2006) of investigated linguistic items within language varieties and compare variety-specific profiles across those language varieties. Behavioral profiles therefore provide form- or sense-specific summaries of the semantic and morpho-syntactic behavior of the linguistic items studied. Based on those profiles, techniques such as hierarchical cluster analysis organize investigated linguistic items by finding (dis)similarities between their profiles across English varieties and by grouping similar varieties together. The benefits of this type of contrastive technique have been shown in a number of studies such as Szmrecsanyi and Kortmann (2011), which involved 25 varieties/languages (11 EFLs, 5 ESLs, 3 standard British English benchmark registers, and 6 European mother-tongue languages) based on part-of-speech classes, Edwards and Laporte (2015) on the use of prepositions in British, American, Singapore, Indian, Hong Kong and Dutch Englishes, Rautionaho et al. (2018) on progressive marking in EFL and ESL, and Deshors (2016) on the *may* vs. *can* lexical alternation in

French- English interlanguage. From a language-learning perspective, this powerful approach can help (i) assess to what extent learners with different native backgrounds can develop L2 varieties that are similar typologically and (ii) capture how different learner populations make use of complex linguistic co-occurrence patterns and ultimately develop abstract (mental) representations of linguistic patterns characteristic of their L2.

Finally, research trend (iii) involves comparing NL and IL for the purpose of predicting and explaining (rather than describing and grouping) learners' linguistic choices. This type of multifactorial regression approach involves capturing systematic co-occurrence patterns of semantic and morpho-syntactic features with linguistic choices in IL. Such approaches help understand when and why English learners make nativelike and non-nativelike linguistic choices, offering a whole new perspective on what it means to compare learner corpora. A landmark within this research trend is the recent development of the MuPDAR protocol (Gries and Deshors 2014, Gries and Adelman 2014) which, still within the confines of the ICM/CIA frameworks, introduces a new procedure to compare learner corpora: Instead of contrasting observed linguistic patterns in NL and IL, data from the native variety are used to predict what a native speaker would have produced in a specific situation, which can then be compared to what a learner actually did say, given each specific linguistic context using a two-set regression approach (described in Chap. 10). As such, MuPDAR is the first real implementation of what Pery-Woodley (1990: 143) wished for nearly 30 years ago: "comparing/contrasting what non-native and native speakers of a language do in a comparable situation", where the "comparable situation" is defined by linguistic and contextual features captured in the multivariate annotation, from which a statistical method predicts what the native speaker would have produced. Such approaches offer the important benefit of being compatible with cognitive-linguistic / usage-based theory, thus informing explorations of the psycholinguistic relevance of corpus findings. Combining technical sophistication and theoretical relevance,

multifactorial approaches to the comparison of learner corpora have started to show how studies primarily based on the over-/underuse of isolated linguistic items do not always tend to do justice to the complexity of interlanguage systems (e.g. see Gries and Wulff (2013) for an application of the approach to the genitive alternation in Chinese- and German-English IL and Gries and Adelman (2014) for an application to subject realization in Japanese conversation by native speakers and learners). Next, we briefly consider the case of Deshors and Gries (2015), a study that brings together a number of the comparison-related issues we have raised so far throughout the present paper.

Deshors and Gries' (2015) study uses the alternation between ditransitive and prepositional dative constructions in native and non-native Englishes as a case in point to illustrate the central (and too often underestimated) issue of comparability of corpus data in LCR. Specifically, the authors examine 1265 occurrences of both constructions across written and spoken corpora: two EFL corpora (French- and German- English IL), three ESL corpora (Hong Kong, Indian, Singapore Englishes) and British English. Their analysis focused on the question of how ESL and EFL speakers' constructional choices differ from those of native speakers of BrE and how those differences are best explored statistically. Nine linguistic predictors of the dative alternation were included in the analysis (including recipient and patient accessibility, semantics, animacy, pronominality, length difference) as was the hierarchical structure of the corpus (with files/speakers nested into variety and into corpus type), something most corpus studies fail to account for. Consequently, the authors are able to show how appropriate mixed-effects modeling techniques can identify fine differences between NS and NNS speakers' behavior while simultaneously controlling for the differences within and between corpora, idiosyncratic effects of speakers and lexical items and the general complexity of IL within a usage-based approach to SLA. Given the complexity of corpora as datasets along with the complexity of linguistic usage patterns in IL in particular, the

adoption of such techniques has become almost inevitable in order to compare learner data rigorously.

5 Future directions

As this chapter shows, the notion of comparability when dealing with learner data is a thorny notion that requires much attention on the part of LCR and SLA analysts to ensure reliable comparisons across varieties, registers, modes, or any other corpus parts. A first main take-home message from our discussion is that ensuring comparability of learner corpora requires much meta- and linguistic data information – particularly speaker-specific information – that is not yet routinely included in corpus compilation and analysis. Much existing corpus work underutilizes the available data, which can lead to distorted views of linguistic patterns characteristic of interlanguage varieties.

A second take-home message is more theoretical in nature and involves the cross-linguistic part of the ICM. Indeed, since the late 1990s, ICM-based studies contrasting interlanguage, target language and learners' native language in single analyses have remained underrepresented compared to CIA-based studies contrasting NL vs. IL and IL vs. IL. In the future, this research gap should be addressed, especially given its potential to explore transfer-related questions in L2 (see Gilquin 2000/2001). While addressing this gap will ensure a more frequent exploitation of the ICM theoretical framework (see Gilquin 2017 for a recent study that begins to address this gap or Paquot (2014) theoretically based on Jarvis's (2000) model, which has gained ground for transfer studies both in LCR and SLA), the resulting more balanced representation of theoretical frameworks will undoubtedly help us connect more than before the fields of LCR and SLA.

Another take-home message involves a seemingly growing disconnect between the methodological framework of ICM/CIA, upon which LCR as a field was created, and the practical implementations that have been developed within the confines of ICM/CIA. As mentioned above, ICM/CIA is set up to contrast language varieties against other language varieties. However, while scholars have slowly begun to make a strong case for the need to account for speaker-specific characteristics, currently these frameworks do not provide a well-developed theoretical apparatus that can account for these characteristics and that are psycholinguistically and SLA-informed. Therefore, there is a gap between current theoretical frameworks, their practical implementations and what empirical studies are showing is needed. While awareness of this disconnect is growing (see Le Bruyn and Paquot forthcoming), much work remains to be done in order to (i) assess, quantify and explore how much between-speaker variability actually diminishes the role of between-variety variability (see Gries 2018) and (ii) how to make sure learner corpora contain enough metadata for each individual speaker (other than speaker proficiency) to do anything with speakers' individual variation.

6 Further readings

Gablasova, D., Brezina V. and McEnery, T. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* 67(S1): 130-154.

This paper revisits the comparative corpus-based method to explore the notions of interspeaker variation in native and non-native language use, the representativeness and comparability of corpus data and how to interpret observed differences across corpora.

Gries, St. Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2): 109-151.

In this paper, Gries explores corpus variability by showing (i) how degrees of variation should be quantified, (ii) how to capture and investigate the source of variation and (iii) how to assess corpus homogeneity based on individual linguistic features. Overall, Gries makes a case for resampling methods and exploratory data analysis accounting for the fact that superficially different results may reflect similar underlying tendencies, the communicative dimensions that surround the use of a given linguistic phenomenon, and actual linguistic phenomena to assess corpus homogeneity rather than relying on word frequencies.

Gries, St. Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*. This paper demonstrates the urgent need to reassess existing quantitative methodological standards in LCR. Specifically, it makes a case for a complete revision of monofactorial over-/underuse approaches to learner corpora by discussing and showcasing the benefits of multifactorial regression-based statistical techniques to research observational (learner) data.

References

Anthony, L. 2010. AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University.

Biber, D. 1988. *Variation Across Speech and Writing*. New York: Cambridge University Press. Bley-Vroman, R. 1983. The comparative fallacy in interlanguage studies: the case of systematicity, *Language Learning* 33: 1-17.

Callies, M., Díez-Bedmar, M.B. and Zaytseva, E. 2014. Using learner corpora for testing and assessing L2 proficiency. In *Measuring L2 proficiency: Perspectives from SLA* (Second Language Acquisition series), P. Leclercq, H. Hilton and A. Edmonds (eds), 71-90. Clevedon: Multilingual Matters.

Callies. M. and Götz, S. 2015. *Learner Corpora in Language Testing and Assessment* (Studies in Corpus Linguistics, Band 70). Amsterdam: Benjamins.

Crosthwaite, P., Lavigne L.Y. C. and Yeonsuk, B. 2016. 'Almost People': A learner corpus account of L2 use and misuse of non-numerical quantification. *Open Linguistics* (2): 317-336.

Deshors, S. C. 2014. A case for a unified treatment of EFL and ESL: A multifactorial approach. *English World-Wide* 35(3): 279-307.

Deshors, S. C. 2016. *Multidimensional perspectives on interlanguage: Exploring* may *and* can *across learner corpora*. Corpora and Language in Use. Presses Universitaires de Louvain.

Deshors, S. C. 2017. Zooming in on verbs in the progressive: A collostructional and correspondence analysis. *Journal of English Linguistics* 45(3): 260-290.

Deshors, S. C. and Gilquin, G. To appear. Modeling World Englishes in the 21st century: New reflections on model-making. In *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, S. C. Deshors (ed). Amsterdam & Philadelphia: John Benjamins.

Divjak, D. S. and Gries, St. Th. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1): 23-60.

Deshors, S. C and Gries, St. Th. 2015. EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1): 130-159.

Edwards, A. 2016. *English in the Netherlands: Functions, forms and attitudes* (Varieties of English around the World, vol. G56). Amsterdam & Philadelphia: John Benjamins.

Edwards, A. and Laporte, S. 2015. Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels. *English World-Wide*, 36(2): 135-169.

Ellis, N. C. and Ferreira-Junior, F. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7: 188-221.

Ellis, N. C., Römer, U. and O'Donnell, M. B. 2016. *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Language Learning Monograph Series. Wiley-Blackwell.

Gablasova, D., Brezina V. and McEnery, T. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* 67(S1): 130-154.

Gilquin, G. 2000/2001. The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3(1): 95-123.

Gilquin, G. 2015. At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared. *English World-Wide* 36(1): 91-124.

Gilquin, G. 2017. A collostruction-based approach to the Integrated Contrastive Model The idiomaticity of causative constructions in English, French and French learner English. Paper presented at the Idiomaticity workshop, University of Oslo, 1-2 Sept 2017.

Gilquin, G. 2018. American and/or British influence on L2 Englishes - Does context tip the scale(s)? In *Modeling World Englishes in the 21st Century: Assessing the Interplay of Emancipation*

and Globalization of ESL Varieties, S. C. Deshors (ed), 187-216. Amsterdam & Philadelphia: John Benjamins.

Gilquin, G. and Paquot, M. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1): 41-61.

Gilquin, G., De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in Contrast. Text-based Cross-linguistic Studies*, K. Aijmer, B. Altenberg and M. Johansson (eds), 37-51. Lund: Lund University Press.

Granger, S. 1998. The computer learner corpus: A versatile new source of data for SLA research. In *Learner English on Computer*, S. Granger (ed), 3-18. London & New York: Longman.

Granger, S. 2002. A bird's eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds.), 3-33. Amsterdam: John Benjamins.

Granger, S. 2004. Computer learner corpus research: Current status and future prospects. In *Applied Corpus Linguistics: A Multidimensional Perspective*, C. Ulla & U. Thomas (eds.), 123-145. Amsterdam: Rodopi.

Granger S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In *Corpora and Language Teaching*, K. Aijmer (ed.), 13-32 Amsterdam: John Benjamins.

Granger, S. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1): 7-24.

Granger, S. and Rayson, P. 1998. Automatic lexical profiling of learner texts. In *Learner English*

on Computer, S. Granger (ed), 119-131. London: Longman.

Granger, S.; Dagneaux, E.; Meunier, F.; Paquot, M. 2002. The *International Corpus of Learner English*. Version 1. Handbook and CD-Rom, Presses Universitaires de Louvain: Louvain-la-Neuve, 2009.

Granger, S.; Dagneaux, E.; Meunier, F.; Paquot, M. 2009. The *International Corpus of Learner English*. Version 2. Handbook and CD-Rom, Presses Universitaires de Louvain: Louvain-la-Neuve, 2009.

Greenbaum, S. and Nelson, G. 1996. The International Corpus of English (ICE) Project. *World Englishes* 15(1): 3–15.

Gries, St. Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 2(1): 276-308.

Gries, St. Th. and Adelman, A. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*, J. Romero-Trillo (ed), 35-54. Cham: Springer.

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

Gries, St. Th. and Wulff, S. 2013. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics* 18(3): 327-356.

Hasselgård, H and Johansson, S. 2011. Learner corpora and contrastive interlanguage analysis. In *A Taste for Corpora.* In *Honour of Sylviane Granger*, F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds), 33-61. Amstedam & Philadelphia: John Benjamins.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Larsen-Freeman, D. 2014. Another step to be taken - Rethinking the end point of the interlanguage continuum. In *Interlanguage. Forty years later*, Z. Han and E. Tarone (eds), 203-220. Amsterdam & Philadelphia: John Benjamins.

Le Bruyn, B. and Paquot, M. Forthcoming. *Learner corpora and second language acquisition research*. Cambridge: Cambridge University Press.

Martinez-Garcia, M. T. and Wulff, S. 2012. Not wrong, yet not quite right: Spanish ESL students' use of gerundial and infinitival complementation. *International Journal of Applied Linguistics* 22(2): 225-244.

McCarthy, M. and Carter, R. 2001. Ten criteria for a spoken grammar. In *New perspectives on grammar teaching in second language classrooms*, E. Hinkle and S. Fotos (eds), 51-75. Mahwah, NJ: Lawrence Erlbaum Associates.

McEnery, T., Brezina, V. and Gablasova. D. (ed.). 2019. Corpus-based approaches to spoken L2 production: Evidence from the Trinity Lancaster Corpus. *International Journal of Learner Corpus Research* 5(2).

McManus, K. 2015. L1-L2 Differences in the Acquisition of Form-Meaning Pairings in a Second Language. *Canadian Modern Language Review* 71(2): 51-77.

Paquot, M. 2007. *EAP vocabulary in EFL learner writing: from extraction to analysis: A phraseology-oriented approach*. Unpublished PhD thesis. Université catholique de Louvain, Centre for English Corpus Linguistics.

Paquot, M. 2014. Cross-linguistic influence and formulaic language: recurrent word sequences in French learner writing. In *EUROSLA Yearbook*, L. Roberts, I. Vedder and J. Hulstijn (eds.), 216-237. Amsterdam: Benjamins.

Paquot, M. and Biber, D. 2015. The impact of genre on EFL learner writing: A MDA perspective. Paper presented at *36th Annual Conference of the International Computer Archive for*

Modern and Medieval English (ICAME 36), *Words, words, words - corpora and lexis*, 27-31 May 2015, Trier University.

Péry-Woodley, M.P. 1990. Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching* 24 (3): 205-214.

Rautionaho, P., Deshors, S. C. and Meriläinen, L. 2018. Revisiting the ENL-ESL-EFL continuum: A multifactorial approach to grammatical aspect in spoken Englishes. *ICAME Journal* 42: 41-78.

Rosen, A. 2018. The fate of linguistic innovations: Jersey English and French learner English compared. In *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, S. C. Deshors (ed.), 171-191. Amsterdam & Philadelphia: John Benjamins.

Scott, M. 2017. WordSmith Tools version 7, Stroud: Lexical Analysis Software.

Selinker, L. 2014. Interlanguage 40 years on. Three themes from here. In *Interlanguage. Forty years later*, Z. Han and E. Tarone (eds), 221- 246. Amsterdam and Philadelphia: John Benjamins.

Szmrecsanyi, B., and Kortmann, B. 2011. Typological Profiling: Learner Englishes versus L2 Varieties of English. In *Second-Language Varieties of English and Learner Englishes: Bridging the Paradigm Gap*, J. Mukherjee and M. Hundt (eds), 167–207. Amsterdam: John Benjamins.

Xiao, R. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28(4): 421-450.

[1] Please note that although the present chapter focuses exclusively on English-language research, the core issues, topics and general principles we discuss also apply to other languages.

[2] See the relatively many studies now that have used Jarvis's 2000 framework to investigate transfer.

[3] In the context of the ICM (Figure 1), CA: Contrastive analysis (in the traditional sense of the term); OL: Original Language; SL: Source Language; TL; Translated Language; NL: Native Language; IL: interlanguage. Outside of Figure 1, however, SL stands for second language, TL stands for target language.

[4] EFL and ESL differ in that with EFL, English is learnt in the environment of one's native language (L1) whereas with ESL, it is learnt in the environment in which it is spoken.

[5] Specific information about ICLE and other widely used UCL corpora can be found here: https://uclouvain.be/en/research institutes/ilc/cecl/corpora.html

[6] Other possible approaches could include Native Language Identification (NLI) studies or Granger's collgrams or studies looking into frequency but not for automatic profiling. Due to space constraints, however, we only focus on the four approaches we list in this paper.