Corpus-linguistic and computational methods for analyzing communicative competence: contributions from usage-based approaches

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

Abstract

This chapter discusses the evolution of the use of corpus-based methods from traditional learner corpus research towards more theory-informed and methodologically more sophisticated second language acquisition work within cognitive/usage-based linguistics. I outline some of the main challenges that usage-based research towards communicative competence is currently trying to overcome, exemplify corpus methods on both a coarse- and a fine-grained level, and offer a few recommendations for current practice. I conclude by discussing desiderata on three levels: data, theory, and statistical analysis.

Key words

usage-based theory, learner corpus research, predictive modeling, MuPDAR(F), frequency, association, dispersion

1 Introduction

The usage-based theory of language (UBTL) is a currently relatively widespread theory that began to emerge in the 1980s. In my view, it emerged first as *Cognitive Linguistics* and/or *Cognitive Grammar* (see esp. Langacker 1987) but as it matured, names such as *exemplar-based* or *usage-based linguistics* became more common; scholars such as Joan Bybee, William Croft, Nick Ellis, and Adele Goldberg are probably best known. The UBTL is based on a variety of assumptions, which in turn have methodological and other implications.

Based on Beckner et al. (2009) and Bybee (2010), we can describe the UBTL as follows. One central assumption is that "the structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes" (Beckner et al. 2009, p. 2), in particular cognitive processes that are domain-general, i.e. not at all unique to language, such as

chunking (which might give rise to constituent structure);
analogy, similarity- and prototype-based categorization processes;
cross-modal association (connections between different sensory modes);
rich memory storage (of exemplars and aspects of the contexts in which they were produced/encountered).

These processes operate over the course of a human's life, which means anyone's mental representation of the encyclopedic, but especially also unconscious linguistic, knowledge changes all the time. As for linguistic structure, there is no *a priori* distinction between different levels of linguistic structure: Just about everything at any level of linguistic analysis – morphemes, words, multi-word units, partially filled expressions (e.g., *you drive me* ADJ), completely schematic syntactic patterns (e.g., NP$_{\text{AGENT}}$ V NP$_{\text{RECIPIENT}}$ NP$_{\text{PATIENT}}$) – is a construction, i.e. a form-function pairing (where *function* includes 'meaning') that is 'frequent enough' (Goldberg 2006, p. 5) or that involves something that is not predictable from its component parts.

The methodological implications from these theoretical assumptions that are relevant in the present context are that

> the sources of data for usage-based grammar are greatly expanded over that of structuralist or generative grammar: Corpus-based studies of either synchrony or diachrony as well as experimental and modeling studies are considered to produce valid data for our understanding of the cognitive representation of language. (Beckner et al. 2009, p.7)

Given these UBTL tenets, corpus data are not just "valid", but particularly valuable to researchers, given that they, depending on the corpus of course, can provide a lot of information about a linguistic expression of interest *E* that the UBTL considers important:

> its frequency on its own (e.g., how frequent is the lemma *give* in a corpus?);
>
> its frequency of co-occurrence with other expressions (e.g., how frequently does *give* occur in the ditransitive? in the prepositional dative? with the meaning of transfer? with other meanings?) And, if *give* occurs in the ditransitive, what other contextual features are likely to be observed (e.g. a human agent, an inanimate patient, an animate recipient, i.e. a transfer scenario?);
>
> the type and token frequencies of other elements where *E* can occur (e.g., how many other verbs occur in the ditransitive (e.g., *tell*, *send*, *show*, *promise*, …) and how frequent it each of them there?);
>
> the degree to which the use of an element changes over the course of time in a longitudinal acquisition corpus (e.g., do children/learners hear *give* most often with the transfer meaning and in the ditransitive? when/how do they extend uses of *give* to other constructions and other functions?);
>
> the degree to which the use of an element changes over time in historical corpora.

Given (i) that corpus data can provide all this information and (ii) that the construct *communicative competence* involves the probabilistic knowledge of which expression to use given a certain context and communicative intention, corpus data provide very useful information for communicative competence in both first and second languages. However, to understand the current state-of-the-art in corpus-based UBTL approaches to learner language, we need to first consider the historical context from, and partially against, which current work has evolved, namely the field of learner corpus research.

2 Historical context

Learner Corpus Research (LCR) is an 'offshoot' of general corpus linguistics focusing largely on the production of non-native speakers (NNS) of some target language; according to Le Bruyn & Paquot (2021, p.1), its origins are outside of the domain of theory-driven (SLA/UBTL) research, and even till now really SLA-driven learner corpus methodology is more of an exception than the rule. Traditional LCR can be seen as having evolved around two central and related methodological frameworks: Contrastive Interlanguage Analysis (CIA) and the Integrated Contrastive Model (ICM); these were largely formulated by Granger (1996) and Gilquin (2000).

The focus of the former is on the exploration of (individual) learner varieties with a focus on English produced by learners (in practice, learners from a variety of mostly European and Asian L1 backgrounds). Sticking with the example of English as the target language (TL), CIA would be concerned with (i) comparing native English (possibly the target of the learner) and the learners' 'variety/version of English', i.e. interlanguage (IL) and (ii) comparisons between different ILs, i.e. the Englishes produced by learners from different L1 backgrounds. The focus of the latter is on cross-linguistic transfer, i.e. the relation between the learner's IL and their L1. In other words, much of LCR focused in L1 transfer errors and deviations from a target-like norm – according to Le Bruyn & Paquot, one of the reasons why LCR has not been popular in SLA. Most recently, the original CIA framework was 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 varieties, but also English as a Second Language varieties and English as a Lingua Franca (see Granger 2015).

Methodologically, it is probably fair to say that LCR as shaped by the above-mentioned analytical frameworks is mostly characterized by two methodological choices: a *linguistic element* to target (most such studies targeted lexical items or certain grammatical constructions) and a *quantitative resolution* to apply to the targeted item and its occurrences in native and learner corpora (most studies involved the notions of over- and underuse, i.e. comparisons of the frequencies often coupled with X^2 and/or log-likelihood/G^2-tests or similar monofactorial or goodness-of-fit tests, see Paquot & Plonsky 2017).[1] Examples of such studies include Altenberg & Granger (2001), Altenberg (2002), Laufer & Waldman (2011), Gilquin & Granger (2011), Gilquin & Lefer (2017), etc.

It is instructive to briefly make a short excursus here and paraphrase the difference between the above kind of LCR work and the kind of UBTL work to be discussed shortly borrowing language from multi-level regression modeling. In multi-level models, we have a response variable, such as test scores of students each taking two tests; the variable with the scores is measured at what is called *level 1*, the *observation level*. But we often also have other variables that we suspect predict the response and that are measured at that same level, *and* we have variables at higher levels, e.g. at the level of, here, the student (level 2), like a student ID, but also variables that describe the student (e.g., BooksAtHome and HoursSelfStudy). But even higher levels are conceivable, as when the students are nested into classrooms (such that each classroom has a different teacher); the classroom then groups all students of one classroom into a group. Such data are exemplified in Table 1 (with parenthesized numbers in the header indicating the level). In such scenarios, the variability of TestScore will partly be due to whatever other level-1 variables one might have, but also to the student-level/level-2 variables BooksAtHome and HoursSelfStudy, and due to level-3 variables, i.e. Classroom.

Table 1: A fictitious multi-level modeling data set

Case	TestScore (1)	StudentID (2)	BooksAtHome (2)	HoursSelfStudy (2)	Classroom (3)
1	11	student1	240	4	a
2	13	student1	240	4	a
3	11	student2	200	3	a
4	9	student2	200	3	a
5	9	student3	160	3.5	b
6	7	student4	160	3.5	b

The point of this excursus is to make it very clear that traditional LCR of the CIA/ICM kind nearly always considered only level-3 predictors and, therefore, was completely or nearly completely acontextual. What traditional over-/underuse LCR studies would have done with Table 1 is the equivalent of computing the means of TestScore (in an LCR study, the mean frequencies of some linguistic element) for each classroom (in an LCR study, for NS and NNS) and done a significance test comparing 11 and 8, while ignoring any other (level-1 or level-2) predictors. Thus, such studies would, here, miss the strong predictive power of BooksAtHome or, to come back to linguistic/SLA contexts, any linguistic/contextual predictor (or even other important level-2 predictors, see Gries 2018 for details). Altenberg & Granger (2001, p.176, Table 2), for example, compute significance tests comparing the frequencies of the lemma MAKE in NNS English of Swedish and French learners to its frequency in NS English.[2]

In other words, while traditional LCR argued in favor of 'comparing/contrasting what NS and NNS of a language do *in a comparable situation*' (Péry-Woodley 1990, p.143, cited by Granger 1996, p.43, my emphasis), such studies did actually not do that, because they did not include any level-1/level-2 predictors that would allow them to state whether the usage situations were comparable; see Gries & Deshors (2014: Sections 1.1, 3) for detailed discussion/exemplification. Much learner language-oriented research in the UBTL paradigm has evolved in response to these shortcomings and, as will become clear, has shifted the focus onto linguistic/contextual level-1 predictors and how their effects differ across L1 backgrounds using multifactorial statistical analysis.

3 Critical issues in current research

The probably most important agenda item (apart from corpus compilation, see Section 6) for corpus-based UBTL approaches to learner language is determining how best to (i) operationalize the cognitive factors recent studies and overviews have proposed are influencing acquisition, processing, and use and then (ii) relate them to central notions of SLA research relevant to communicative competence (e.g., complexity, accuracy, fluency).

As for the former, the measure simplest to operationalize – *token frequency* – is also one whose importance, while long taken for granted, might be less obvious than has long been assumed (see Section 6.2). No one denies that *association* plays a role for learning, but how do we measure it best using corpus data (see Section 4 for some discussion) for each situation in which it is relevant? What corpus data do we include in our corpus measures to infer degrees of *prototypicality*? How do we operationalize *salience* in discourse? What is the best way to tackle *dispersion* in a corpus? For many of these notions we have reasonable proxies – see Ellis, Römer,

& O'Donnell 2016 for one of the most well-rounded (book-length) studies providing corpus-linguistic approximations for many of the above terms – but these are issues that every current UBTL study needs to address in one way or the other. This is especially so because of the nature of the UBTL itself: A theory that makes rich memory storage *and* domain-general learning mechanisms its default assumptions certainly seems appealingly 'big-picture' and unifying, but with those starting assumptions also comes incredible complexity. This is in contrast with, say, more modular theories because if modularity is the default assumption, one is not automatically under the pressure of (Lakoff 1990, p.40) Cognitive Commitment to "[provide] a characterization of general principles for language that accords with what is known about the mind and brain from other disciplines" – one is freer to choose that a certain postulated mechanism is specific to the language module.

As for the latter, while complexity and fluency can be addressed fairly well using even automated measures, accuracy is different: Typically, it requires laborious hand-coding and leads to varying degrees of reliability, but Polio & Yoon (2021) introduce a number of automated measures of accuracy drawn from usage-based theories of SLA by measuring how likely bi- and trigrams occurring in learner texts are to also co-occur in large NS reference corpora. Data from three learner corpora show that bi- and trigrams not occurring in reference corpora are in general considered erroneous by human judges. They also find that their automated measures of accuracy typically pattern with classical accuracy measures – as desired – and not with complexity measures. Finally, their results show that up to half of the variance of hand-coded error counts is accounted for by their automated measures.

Taken together, the maybe most fundamental challenge will be to determine how the extremely high-dimensional exemplar space that the UBTL postulates can be 'modeled' using usage data and how cognitively realistic this 'model' will or should be. I personally believe that one reason for why over time the moniker *usage-based linguistics* overtook *cognitive linguistics* is that cognitive linguists realized that much of their work was not cognitive in the cognitive-science kind of sense, but usage-based, but even 'just' being usage-based requires juggling many dimensions of information and being even just 'somewhat' cognitively realistic requires doing so while keeping these dimensions separate rather than conflating them into easy-to-use but cognitively unrealistic indices. Gries (2019), for instance, shows how even a simple method such as collostructional analysis, which quantifies the association of a word to a construction with a single value (see next section), would need to become much more precise by breaking this one value up into at least four or five different dimensions, and similar challenges abound for probably all corpus-linguistic operationalizations of cognitive mechanisms.

4 Main research methods

The main research methods within corpus-based UBTL work do not really differ much from the corpus-linguistic methods one finds in any (sub-)discipline, because, frankly, there are not many fundamentally different corpus-linguistic methods and all of them are ultimately derivatives of frequencies of (co-)occurrence. It seems appropriate, in fact, to view corpus-linguistic methods performed on an existing corpus as a combination of (i) a very small number of *retrieval operations* of some element *E* from a corpus (part, such as a file, a register, …) followed by (ii) one or more of a larger number of *statistical operations* performed on/with the retrieved element(s).

4.1 Level of resolution 1: a slot (in a construction in a corpus (part))

At one level of resolution, the retrieval operation involves retrieving one or more linguistic element(s) $E_{1\text{-}n}$ from a corpus (part) and either providing its/their frequency/ies in general (often normalized to per million words) or providing its/their frequency/ies in/with something else; as mentioned in Section 1, examples include how often is *give* used in a corpus (part) and/or how often is it used with a certain meaning or in a certain grammatical construction, in which case the normalized frequency becomes a conditional probability, as in $p(give|\text{ditransitive})$.

While conditional probabilities are often used as the simplest of association measures (AMs), corpus linguists have now for decades preferred to express the association between an element *E* (e.g., *give*) and some other element *X* (e.g., the ditransitive) not just with conditional probabilities, but with association measures. Consider Table 2 for a schematic 2×2 co-occurrence frequency table of the type that is widely used in corpus-linguistic studies (cognitive or otherwise); there, the element *E* of interest in the upper row might be a word (e.g., *give*) and the co-occurring element *X* might be a construction (e.g., the ditransitive). Thus,

the row total $a+b$ would be the frequency of the *give* in a corpus;
the column total $a+c$ would be the frequency of the ditransitive in a corpus;
the cell *a* would be the frequency of *give* in the ditransitive.

Table 2: A schematic 2×2 co-occurrence frequency table

	Co-occurring element *X*	Other elements (not *X*)	Totals
Element *E*	*a*	*b*	*a*+*b*
Other elements (not *E*)	*c*	*d*	*c*+*d*
Totals	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*

However, quantifying the co-occurrence of *give* in the ditransitive with the conditional probability $^{a}/_{a+b}$ or $^{a}/_{a+c}$ neglects what happens in the other row (with $^{c}/_{c+d}$) or the other column ($^{b}/_{b+d}$) – most AMs therefore use more of the information in Table 2 and the most frequent AMs – the log-likelihood value G^2, (log) odds ratio (OR), pointwise *MI*, *t*, *z*, conditional probability $p(E|X)$, and ΔP) – are all derivable from one and the same statistical approach (logistic regression), yet still behave differently. Some

reflect association (like the odds ratio or ΔP) while some also reflect the frequency of the element(s) in question (e.g., G^2 or *t*);
consider the row/column of Table 2 containing cell *a* whereas others also consider more information in the table (the other row/column or the column/row totals);
return a measure of mutual/*bi*directional association between *E* and *X* whereas others are *uni*directional and, thus, distinguish the direction of association $E \rightarrow X$ from $E \rightarrow X$.

Studies that focused on learner collocations of lexical items and/or phraseologisms have often used *MI* (an AM reflecting bidirectional association and, thus, often returning very rare collocations/phraseologisms) or *t* (an AM reflecting bidirectional association and frequency and, thus, often returning frequent items). For example, Paquot et al. (2021) assess phraseological complexity by checking how much words constituting word combinations in a syntactic dependency relation in learner texts are attracted to each other in a large reference corpus

(ENCOW14 AX). They use mean *MI*-scores as a measure of phraseological complexity and correlate those with time (in a longitudinal corpus) and external/independent Oxford Quick Placement Test scores. They show that time and institutional training do not correlate with the development of phraseological complexity per se – language proficiency and external test score changes from one year to the next matter more.

Studies that focused on the above kind of example – co-occurrence of words with constructions – usually adopted an AM that has been widely used in *collostructional analysis* (a family of methods to explore the co-occurrence preferences of words and/in constructions), namely the *p*-value of a Fisher-Yates exact test (see Stefanowitsch & Gries 2003, Gries & Stefanowitsch 2004a, b). For example, Wulff & Gries (2011) reject a binary notion of accuracy and argue, as I did here at the end of Section 1, that

> accurate mastery of a language entails the acquisition of constructions at different levels of complexity and schematization, as well as *knowledge of the probabilistic tendencies underlying their target-like combination* (p. 63, my emphasis)

They show that the verbs that NS and NNS prefer to use in the two constructions of the dative alternation are highly similar, sometimes to the point that the NNS uses are *more* in line with linguistic theory than what NS do (e.g., the learners' strong preference to use *send* in the prepositional dative); they also report results that are, on the whole, similar for the *to* vs. *ing*-complementation alternation (*I prefer swimming* vs. *I prefer to swim*). They conclude that 'learners have constructions' and that (p. 81f.)

> accuracy will increase proportionally to the extent that learners succeed in making the right generalizations regarding which form […] is mapped onto which function […]. Note that "making the right generalizations" amounts to nothing else than learners being able to extract prior probabilities (e.g., the knowledge that *give* is more frequent than *donate*) as well as posterior/conditional probabilities (e.g., the knowledge that *give* is used ditransitively more often than donate) from the multidimensional input space.

While these kinds of association-based approaches are useful, they are still limited because, if the association scores are not used for any subsequent analysis, these approaches do not involve many UBTL predictors or features. Thus, their predictive power for actual linguistic choices is by definition moderate – the approaches discussed next change this considerably.

4.2 Level of resolution 2: a specific linguistic choice in a concordance line

At this level of resolution, the retrieval operation typically involves retrieving one or more linguistic element(s) $E_{1(-n)}$ from a corpus (part) together with their contexts; a less frequent yet still important alternative is to retrieve contexts in which *E* could have been used but wasn't. An example of a hybrid strategy involving both kinds of retrieval could be used to study *that*-complementizer realization/omission (e.g., *I know that/- everyone loves Babylon 5*): One might (i) retrieve all instances of *that*, (ii) identify which of them are examples of object complementation like the above, (iii) retrieve all forms of all main-clause verbs ever used with *that* to, finally, (iv) identify which of them are examples that do not have a complementizer but could have one.

All resulting hits could then be annotated for whatever predictors/features seem relevant to explain the response, *E*'s form or its presence/absence. Crucially and in contrast to much traditional LCR, such features include level-1 features varying from case to case *and* higher-level features (e.g., speakers producing multiple examples or words that are observed with choices, e.g., the main-clause verb *thought* in the above example). This annotation can then be used with statistical predictive-modeling tools such as regression or tree-based approaches.

For example, Gries & Wulff (2013) model the genitive alternation (*of* vs. *s*) from a UBTL perspective. They annotate approximately three thousand matches from NS of English and Chinese and German NNS of English for 12 predictors from various levels of linguistic analysis – phonology, morphosyntax, semantics, and psycholinguistics – and fit a regression model with all predictors. Crucially, they permit each predictor to interact with L1 (Chinese vs. German vs. English/native) to determine whether the factors that govern NNS choices are significantly different from those that govern NS choices. They obtain a significant (p<10^{-200}) and excellent model fit (C=0.96) and find, among other things, that NS and NNS differ in terms of the effects of the genitive construction's semantics and the specificity of the possessor and the possessum. They then discuss how their multifactorial approach involving 12 predictors differs from what would be the traditional chi-squared test LCR approach that would feature one predictor at a time.

An extension of this regression approach is the recently-developed MuPDAR(F) protocol (for Multifactorial Prediction and Deviation Analysis using Regression/(Random Forests), see Gries & Adelman 2014, Gries & Deshors 2014, 2020):

> one applies a model/classifier to the part of the data covering the reference speakers (RS, in LCR contexts, the NS);
>
> if that first model/classifier works well enough, it is used to impute for each situation a target speaker (TS, in LCR contexts, the NNS) was in what the RS would have said in the exact same linguistic context;
>
> then, one determines how the actual TS choices relate to the imputed ones: how much, if at all, does the TS choice deviate from the imputed RS choice?
>
> finally, one explores what explains those TS choices that are unlike the imputed ones with second model/classifier.

MuPDAR(F) has led to many interesting results in studies such as Deshors & Gries (2016) and Kolbe-Hanna & Baldus (2018) on *ing* vs. *to*-complements, Wulff & Gries (2015, 2019, 2021) on prenominal adjective order, particle placement, and genitives respectively, Kruger & De Sutter (2018), Gries & Wulff (2021) on adverbial clause ordering, Schweinberger (2020) on adjective amplification, and others. For illustration, I will discuss Lester (2019) who studies the realization/omission of *that* as a relativizer (e.g., *Bester hated the way that/- telepaths were treated*). Eight hundred relative clauses with/without *that* (40% of those from native speakers, the remainder from German and Spanish learners) were retrieved from two corpora and annotated for 13 variables (including what would normally be the response variable, i.e., *that*-realization) including task type, semantic predictors, structural/complexity predictors, priming and disfluencies.

He then fits a generalized additive mixed model (GAMM) on the native speaker data, cross-validates it with a bootstrap, applies it to the learner data, and computes how much the learner choices deviate from the imputed NS choices, which became the response variable in a second GAMM. That model results in several significant linear and non-linear predictors. To give a few examples of the findings, the Spanish learners perform in a more nativelike fashion than the

German ones, all learners overuse *that* for subject, predicate-nominal, and direct-object roles of the relative-clause heads, and self-priming effects differed between the German and the Spanish learners. More generally, the data do not support the study's initial expectation that NNS would follow the same processing-based strategy (of producing *that* in complex contexts) – instead, learners underproduce *that* in structurally complex contexts and under production difficulty. This study is a great example of how applying advanced statistical methods to offline observational data can still shed light even on the kind of online processing-related/cognitive differences and strategies between NS and NNS that give rise to differences in the degree of attainment of communicative competence.

5 Recommendations for practice

There are actually few recommendations specific to corpus-linguistic UBTL research of communicative competence – the following pertains to just about all corpus-linguistic studies.

On the retrieval level, it is obviously important to use search expressions that maximize recall of the target element *E*, but proper context retrieval and sampling is nearly as important. As for the former, one often needs considerably more context of each instance of *E* than suspected to annotate especially semantic, discourse-functional, or psycholinguistic predictors: Annotating discourse givenness, inferrability, or priming requires at least several sentences of context. Also, too many studies are still sampling on the level of the individual data point – retrieving all instances of *E* and then taking a random sample of them – when that is sub-optimal. To

achieve better/decent numbers of data points for random effects;
allow for proper consideration of priming effects;
be able to account for, say, within-conversation learning/habituation effects,

one should sample on the level of the speaker/conversation.

On the statistical level, the importance of thorough (i) exploration of the data and (ii) diagnosis and validation of one's model cannot possibly be overstated. The variables in the data need to be checked for data entry errors and consistency, outliers, and the need for conflation, general distributional characteristics (maybe requiring transformations); models need to be checked for collinearity, cases with huge leverage, their residuals, overdispersion, maybe validation, etc. In the online supplement to Gries (2021), approximately 70% of the input/output in this one modeling application are concerned with exploration, diagnostics, etc. – these kinds of things are not nice-to-have add-ons; they are obligatory! Finally, reporting of methods and results usually needs to be more comprehensive, to ensure replicability, but also to allow readers to evaluate results better. For instance, there simply is no good reason not to report overall model statistics (significance tests, but also R^2s), but these are still often not provided. However, the field has improved considerably in these regards.

6 Future directions

6.1 On the data side of things

If we were allowed to move the field forward with only a single thing, it would have to be 'more and better corpus compilation', and I am saying this as someone who has only been involved in two small corpus compilation projects myself. Essentially, we need more of 'everything':

> more coverage of more L1s and L1-L2 configurations, more diverse registers/genres, and more proficiency levels, and, importantly, more input corpora;
>
> more longitudinal and more multilingual corpora;
>
> more annotation on characteristics of the speakers such as proficiency levels, learners' L1s and other background characteristics (age, amount of previous instruction in hours (not years), country of residence, SES/parental education, cognitive variables such as motivation information, results from aptitude tests, working memory capacity, etc.), and characteristics of the context of learning (naturalistic? instructed?);
>
> more information about the speakers' creation of written data (e.g. from screen-casting and key-logging tools as used in translation research).

Le Bruyn & Paquot (2021) and Tracy-Ventura, Paquot, & Myles (2021) indicate that many more diverse corpora are now being compiled, but we still have a long way to go before we can do all kinds of analyses of communicative competence UBTL researchers are interested in.

In addition, in order for UBTL researchers to be able to '(more) easily' identify the frequencies of co-occurrence that so much in UBTL hinges on, we need good assessments of how well recent high-powered automatic NLP tools (e.g. tagging/parsing R packages such as `NLP`/`openNLP` or `udpipe` or Python-based tools such as `spacy` [https://spacy.io]) and many others work on learner data (see Meurers & Dickonson 2017 or Kyle 2021 for overviews of the use of NLP technologies in SLA/LCR). Relatedly, the field needs better ways of dealing with formulaic/prefabricated language, multi-word units, and phraseologisms. While there is a general recognition that these are important concepts, their definition/measurement, their acquisition, and the implications they would have for both communicative competence (in certain contexts) and theory development require much more and rigorous empirical work.

6.2 On the theoretical side of things

While the notion of frequency has been at the forefront of the UBTL, it is neither the only important notion let alone the most important one – many other distributional characteristics are just as essential or even more so (even if they, technically, of course derive from frequency data).

One important notion that is as widely neglected in even the best scholars' studies as it is actually easy to measure is *dispersion* (i.e., the degree to which words/constructions are evenly distributed in a corpus). Most scholars do not use it because they do not know about it or because they think that dispersion is so highly correlated with frequency that it is unnecessary. However, this correlation breaks down in exactly the range of frequencies that are of most interest to most linguistic studies (intermediately frequent content words, Gries 2020), which means that scholars who think they are controlling for frequency effects are likely not doing so (well). Second, there are also studies showing that dispersion can have higher predictive power than frequency (Baayen 2010, Gries 2010) and might therefore be a better measure of 'commonness'. Third, dispersion is straightforwardly integratable into UBTL/SLA research via the notions of recency and (associative) learning theory so future research trying to be cognitively realistic would do well to include it.

Another set of dimensions of information UBTL needs to attend to more involves several broader and often *information-theoretic ways of including co-occurrence information* that speakers seem to unconsciously keep track of. For example, McDonald & Shillcock (2001) show that the degree to which a word influences the frequencies of its collocates is more predictive of reaction times than frequency; for instance, Berger et al. (2017) study how much measures pertaining to words' contexts (including relative entropy and measures based on association tasks) are correlated with human ratings of lexical proficiency. Linzen & Jaeger (2015) find that the entropy reduction of potential parse completions is correlated with reading times of sentences involving the DO/SC alternation; e.g., *accept* in *Garibaldi accepted Sinclair was right* has a lower entropy of possible complementation patterns compared to *forgot* in *Garibaldi forgot Sinclair was right*, which is reflected in reading speeds. Blumenthal-Dramé (2016, p. 500) reports that the entropy of verbs' subcategorization frames correlates with activity in the anterior temporal lobe 200-300 ms after the stimulus. And, Lester & Moscoso del Prado (2017) find that entropies of syntactic distributions affect response times of nouns in isolation and their ordering in coordinate NPs and arrive at the Construction Grammar par excellence conclusion that

> words are finely articulated syntactic entities whose history of use partially determines how efficiently they are produced […] Perhaps words and syntactic structures are much more tightly linked than is typically acknowledged.

Thus, while frequency is often a good first explanatory step and frequencies underlie virtually all more refined measures, subsequent analysis will ultimately have to face that the exemplar-space kind of knowledge the UBTL assumes will require a much broader perspective.

6.3 On the statistical side of things

One of the main developments has already begun: the move away from over-/underuse of frequencies aggregated over many speakers etc. and without level-1 predictors of *E*. Gries (2018) re-analyses an older study, showing that such studies are uninformative at best (and misleading at worst) because they ignore nearly everything but L1; therefore, when these studies are replicated, their predictive power is close to zero as is their relevance to theoretical work (see Tracy-Ventura et al. 2021, p. 420f. for similar views). The move towards modeling is therefore good news because it allows researchers to paint a more comprehensive picture of *E*'s acquisition, learning, and use.

With that greater comprehensiveness come greater challenges. Counter to widespread belief, proper (predictive) modeling is not a simple endeavor, especially given the complexity of the questions being studied. We need to

> include (many) linguistic/contextual level-1 predictors of *E* and we need to allow numeric predictors to be curved – few cognitive processes follow a straight line – and predictors of interest to participate in relevant interactions;
> include higher-level predictors/random effects regarding L1, circumstances of production (genre, topic, etc.), and speaker-specific effects plus proper follow-up analyses of such effects that, currently, are very rare.

Recent relevant examples for such studies are Verspoor, Lowie, & Wieling (2021), who discuss individual differences and non-linearity in their study of learner performance or Gries's (2021) (didactic) methods showcase of a corpus-based LCR/SLA study of *that*-complementation. In addition, modeling techniques like structural equation modeling would be a useful next step to better handle the interplay of many important and often intercorrelated (and, thus, redundant) variables.

Finally, I would like to see a greater reliance also on exploratory statistics – either as a preparatory tool before modeling applications or to really just explore data. Specifically, the following three techniques hold promise:

variants of cluster analysis such as *fuzzy clustering*, which, unlike traditional clustering tools permit elements to be a member of more than one cluster and thus can do better justice to the overlapping nature of natural-language categories;

social network analysis, which is used to great effect by Ellis, Römer, & O'Donnell (2016) to identify groups of verbs in constructions on the basis of exactly the kind of distributional behavior that UBL considers essential to language acquisition and learning;

association rules, i.e. an exhaustive algorithm to identify predictive if-then statements within large amounts of categorical data of the type that result from annotating corpus data for qualitative/categorical variables.

Once we have more and better data, such statistical advances will permit LCR practitioners to leave behind their monofactorial past and be of much more relevance to SLA and UBL theorists.

7 Discussion questions

Why is it so crucial to include multi-word units/phraseologisms (more) in studies of learners' communicative competence?

How can one operationalize corpus-linguistically notions such as productivity, prototypicality, salience, or surprisal?

How can we use mixed-effects modeling approaches for research on individual variation?

8 Suggestions for further reading

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

Gries, St. Th. & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora 9*(1), 109-136.

Le Bruyn, B. & Paquot, M. (eds.). (2021). *Learner corpora and second language acquisition research*. Cambridge: C.U.P.

Notes

[1] While the introduction of CIA2 led to the terminological replacement of *over-/underuse* by *over- and underrepresentations*, this terminological change had no substantive theoretical implications.

[2] The reported statistics are non-replicable because Table 1 in that paper misrepresents the size of one of the learner corpora by approximately a factor of 10.

References

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 37-54). Amsterdam & Philadelphia: John Benjamins.

Altenberg, B. & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics 22*(2), 173-195.

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

Beckner, C., Blythe, R., Bybee, J. Christiansen, M. H., Croft, W. Ellis, N. C., Holland, J., Ke, J. Larsen-Freeman, T., & Schoenemann, T.. (2009). Language is a complex adaptive system. *Language Learning 59*(S1), 1-26.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Educational Technology & Society 20*(2), 201-212.

Blumenthal-Dramé, A. (2016). What corpus-based Cognitive Linguistics can and cannot expect from neurolinguistics. *Cognitive Linguistics 27*(4), 493–505.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: C.U.P.

Collins, P. (2009). *Modals and quasi modals in English*. Amsterdam: Rodopi.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: [...]*. Language Learning Monograph Series. Wiley-Blackwell.

Gilquin, G. (2000). The integrated contrastive model: Spicing up your data. *Languages in Contrast 3*(1), 95-123.

Gilquin, G & S. Granger. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 55-78). Amsterdam & Philadelphia: John Benjamins.

Gilquin, G. & M.-A. Lefer. (2017). Exploring word-formation in Learner Corpus Research: A case study on English negative affixes. Paper presented at the Learner Corpus Research conference 2017, Bolzano, Italy.

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

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

Gries, St. Th. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In St. Th. Gries, S. Wulff, & M. Davies (eds.), *Corpus linguistic applications: current studies, new directions* (pp. 197-212). Amsterdam: Rodopi.

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

Gries, St. Th. (2019). *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Leiden & Boston: Brill.

Gries, St. Th. (2020). Analyzing dispersion. In M. Paquot & St. Th. Gries (eds.), *A practical handbook of corpus linguistics*. Berlin & New York: Springer.

Gries, St. Th. (2021). (Generalized linear) Mixed-effects modeling: a learner corpus example. *Language Learning*.

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

Gries, St. Th. & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora 9*(1), 109-136.

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

Gries, St. Th. & Stefanowitsch, A. (2004b). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (eds.), *Language, culture, and mind* (pp. 225-236). Stanford, CA: CSLI.

Gries, St. Th. & 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.

Gries, St. Th. & Wulff. S. (2021). Examining individual variation in learner production data: a few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied Psycholinguistics 42*(2), 279-299.

Kolbe-Hanna, D. & Baldus, L. (2018). The choice between *-ing* and *to* complement clauses in English as first, second and foreign language. Paper at ICAME 39, University of Tampere, Finland.

Kyle, K. (2021). Natural language processing for learner corpus research (introduction to the special issue). *International Journal of Learner Corpus Research 7*(1), 1-16.

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics 1*(1), 39-74.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites.* Stanford: Stanford University Press.

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning 61*(2), 647-672.

Le Bruyn, B. & Paquot, M. (2021). Learner Corpus Research and Second Language Acquisition: an attempt at bridging the gap. In B. Le Bruyn & M. Paquot (eds.), *Learner corpora and second language acquisition research* (pp. 122-147). Cambridge: C.U.P.

Lester, N. A. (2019). *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research 5*(1), 1-32.

Lester, N. A. & Moscoso del Prado Martín, F. (2017). Syntactic flexibility in the noun: evidence from picture naming. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2585-2590.

Linzen, T. & Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: evidence From subcategorization distributions. *Cognitive Science 40*(6), 1382-1411.

McDonald, S. A. & Shillcock, R. C. (2001). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech 44*(3), 295-323.

Meurers, D. & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning 67*(S1), 66-95.

Paquot, M., Naets, H., & Gries, St. Th. (2021). Using syntactic co-occurrences to trace phraseological development in learner writing: verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (eds.), *Learner corpora and second language acquisition research* (pp. 122-147). Cambridge: C.U.P.

Schweinberger, M. (2020). A corpus-based analysis of differences in the use of very for adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research 6*(2), 163-192.

Stefanowitsch, A. & Gries, St. Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics 8*(2), 209-243.

Tracy-Ventura, N., Paquot, M.n & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (eds.), *The Routledge Handbook of SLA and Corpora*, 411-426. New York & London: Routledge.

Verspoor, M., Lowie, W., & Wieling, M. (2021). L2 developmental measures from a dynamic perspective. In B. Le Bruyn & M. Paquot (eds.), *Learner corpora and second language acquisition research* (pp. 172-190). Cambridge: C.U.P.

Wulff, S. & Gries, St. Th. (2015). Prenominal adjective order preferences in Chinese and German L2 English: a multifactorial corpus study. *Linguistic Approaches to Bilingualism 5*(1), 122-150.

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

Wulff, S. & Gries, St. Th. (2021). Explaining individual variation in learner corpus research: some methodological suggestions. In B. Le Bruyn & M. Paquot (eds.), *Learner corpora and second language acquisition research* (pp. 191-213). Cambridge: C.U.P.