Language Teaching (2022), **55**, 233–244
doi:10.1017/S0261444821000392

CAMBRIDGE UNIVERSITY PRESS

PLENARY SPEECH

Applied corpus linguistics for language acquisition, pedagogy, and beyond[1]

Ute Römer*

Georgia State University, Atlanta, USA
*Corresponding author: Email: uroemer@gsu.edu

Abstract

This plenary speech provides an overview of applications of corpus research in several core areas of applied linguistics, including second language acquisition and language assessment. It does this by showcasing a number of recent studies carried out by or with involvement of the author. These studies all focus on phraseological aspects of language and demonstrate the importance of studying its patterned nature. The studies also illustrate how corpora and corpus-analytic techniques can allow us as applied linguists to contribute to solving problems in other disciplines (such as legal scholarship or music theory) and hope to thereby encourage more interdisciplinary research collaborations between corpus linguists and scholars from other fields.

1. Introduction

My main goal for this paper is to demonstrate that Corpus Linguistics (CL) has a lot to offer to the Applied Linguistics (AL) community. I will provide an overview of some recent work I have done at the intersection of CL and AL and in collaboration with colleagues from different sub-fields of AL and beyond. I will introduce these colleagues to you along the way. The studies I am going to share are grouped into three topics: usage-based second language acquisition (SLA), corpus-based language assessment, and empirical legal language analysis. What ties these studies together is the use of corpora as well as a shared focus on phraseology, or the study of language patterns, which recognizes the inseparability of lexis and grammar (Römer, 2009) and treats the two in an integrated fashion.

Since the 1970s, corpora (i.e., large, systematic, computerized collections of texts) and corpus-analytic techniques have enabled insights into the nature of language that no other types of data or methodology had before. Corpora have profoundly changed the ways in which we see and describe language(s). As John Sinclair, a leading first-generation corpus linguist, who introduced me and other scholars around the world to this new way of doing language analysis, observed in his seminal monograph *Corpus, Concordance, Collocation*, 'the language looks rather different when you look at a lot of if at once' (Sinclair, 1991, p. 100). The power, according to Sinclair, lies in observing what he calls 'repeated events' (Sinclair, 2004, p. 28). Over the last half century, not only general theoretical linguistics but also a range of areas within AL have benefitted from this new, different perception of and perspective on language. This last point is illustrated well by the program for this conference. Treating the AAAL 2021 program as a corpus, I looked up 'corpus' and 'corpora' to see how many papers discussed or mentioned these words in their abstracts. I found that within the 'on demand' section of the program, there are 158 records that use the word 'corpus' and 140 that contain 'corpora'. To see whether these instances of corpus/corpora were well dispersed across AAAL's thematic strands, I looked at each abstract they occurred in and found that, out of the 22 strands, 20

[1]This paper is based on a plenary talk delivered by the author at the American Association for Applied Linguistics (AAAL) 2021 Virtual Conference on 21 March 2021.

© The Author(s), 2021. Published by Cambridge University Press

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

include papers that talk about 'corpora' or 'corpus'. The papers within these strands cover a wide range of topics and fields of application, ranging from language policy (e.g., the paper by Peter J. Torres) to language assessment (e.g., the paper by Geoffrey LaFlair and Jesse Egbert) to languages for specific purposes such as aviation English (e.g., the paper by Malila Prado and Noriko Ishihara).

2. Corpora and phraseology

As I hinted at earlier, a particular strength of CL lies in allowing us to uncover the patterned nature of language. In a 2008 article on evaluation in academic writing, I used traditional corpus analytic techniques like keyword and word frequency analyses to automatically extract items of evaluative meaning from a corpus of book reviews but found that those steps were not successful, at which point I concluded, echoing what Sinclair and others have pointed out too, that '[t]he word is not enough' (Römer, 2008, p. 121). A phraseological approach, however, was successful and allowed me to automatically and semi-automatically extract units of meaning from the corpus. Again referring back to Sinclair, language has a 'phraseological tendency' (Sinclair, 2004, p. 29, 1991): words do not appear in isolation, but go together and make meanings by their combinations. Sinclair, inspired by scholars like Ludwig Wittgenstein and John Rupert Firth, also reminds us that 'the normal primary carrier of meaning is the phrase and not the word' (2008, p. 409). Words in isolation do not mean anything on their own but start to take on meaning(s) when we put them in context.

Another thing we learn from corpus researchers, and I am here quoting Michael Stubbs, is that '[m]any words are frequent because they are used in frequent phrases' (Stubbs, 2009, p. 119). Let us look at the word 'thing' to illustrate this last point. A corpus can tell us why 'thing' is the most frequent noun in English. Table 1 lists the 20 most frequent word clusters of different lengths that contain 'thing' in an 80-million word subset of academic prose of the Corpus of Contemporary American English (COCA). We see that 'thing' is highly frequent because it occurs in frequent phrases that have important communicative functions and that we use to create meaning in language; for example, 'the same thing', 'a good thing', and 'there is no such thing'. With its focus on language patterns, corpus research has also taught us that when a word occurs in different patterns, it also tends to express different meanings. Examples (1) and (2) are taken from the *Collins COBUILD English Dictionary* (discussed in Hunston & Francis, 1998, p. 51), and show that when the verb 'recover' is followed by 'from', followed by a noun or noun phrase, it means 'to get better'. When 'recover' is immediately followed by a noun or noun phrase, it means 'to get back'. This demonstrates that linguistic usage patterns and meanings are 'mutually dependent' (Hunston & Francis, 2000, p. 3) and cannot be separated.

(1) *He is recovering from a knee injury.* (pattern: 'recover' + 'from' + N; meaning: 'get better')
(2) *Police … recovered stolen goods.* (pattern: 'recover' + N; meaning: 'get back')

The phraseological concepts used in the case studies I am about to discuss include *n-gram*, *p-frame*, and *verb-argument construction*. N-grams are fixed sequences of *n* words; for example, 'I don't know' is a 3-gram and 'as soon as possible' a 4-gram. P-frames or phrase-frames are n-grams with one or more variable slot(s) that can be filled with so-called 'variants'. So, in the p-frame 'as * as possible', the * slot can be filled with multiple words, including 'soon', 'much', and 'many'. Figure 1 shows the p-frame 'I don't * if' with frequency information for its five variants in MICASE, the Michigan Corpus of Academic Spoken English. Within the theoretical framework of Construction Grammar, constructions have been defined as fixed pairings of form and meaning that are entrenched in the speaker's mind and conventionalized in the speech community (Bybee, 2010; Goldberg, 1995). Verb-argument constructions (VACs) are a type of construction consisting of a verb and its grammatical arguments; for example, 'V N' (transitive construction) or 'V *in* N' (verb plus 'in' plus noun or noun phrase).

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

Table 1. Top-20 'thing' clusters (3–6 words long) in COCA-academic (list generated with *AntConc*; Anthony, 2019)

Rank	Frequency	Cluster
1	710	the same thing
2	335	the only thing
3	334	for one thing
4	272	thing to do
5	267	a good thing
6	259	the first thing
7	251	no such thing
8	234	the right thing
9	217	is one thing
10	214	such thing as
11	204	no such thing as
12	197	such a thing
13	159	is no such thing
14	154	there is no such thing
15	144	most important thing
16	143	one thing to
17	142	is no such thing as
18	141	one thing that
19	139	there is no such thing as
20	136	the whole thing

```
I don't * if                276        5
I don't know if             256
I don't care if              10
I don't remember if           6
I don't think if              2
I don't mind if               1
```

Figure 1. The phrase-frame 'I don't * if' with its five variants in MICASE (results generated with *kfNgram*; Fletcher, 2007).

3. Phraseology and applied CL

Let me now take you on a journey through three types of case studies, all of them illustrating phraseology research based on corpora.

3.1. Usage-based SLA

I will start by highlighting three recent corpus studies that, I think, all contribute to a better understanding of SLA and have implications for language pedagogy. My co-author on the first study, published as Römer and Yilmaz (2019), is Selahattin Yilmaz. The study focuses on a set of 19 VACs in advanced Turkish learner writing in English. The aim of this study was to find out in what ways learners' constructional knowledge is influenced by English usage and by cross-linguistic transfer from their first language (L1). We used data from two corpora, the Turkish sub-component of the International Corpus of Learner English (ICLE) and the British National Corpus (BNC, representing L1 English usage) on a selection of VACs. The selected VACs all consisted of a verb followed by a

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

Table 2. Top-10 verbs in 'V *about* N' (table adapted from Römer & Yilmaz, 2019, p. 119)

	BNC		ICLE Turkish	
Rank	Verb	Frequency	Verb	Frequency
1	*talk*	3,832	*think*	51
2	*think*	3,153	*talk*	35
3	*be*	2,827	*be*	21
4	*know*	1,812	mention	15
5	worry	910	care	13
6	say	721	learn	12
7	bring	712	complain	5
8	hear	604	discuss	5
9	forget	556	do	5
10	write	517	*know*	4

preposition followed by a noun or noun phrase (e.g., 'V *with* N' exemplified in 'I like to go with the flow'). We found that advanced L1 Turkish learners are sensitive to verb-in-VAC distributions in L1 usage and that their verb-VAC associations, that is the connections between lexical items and constructions, overlap with those of L1 English speakers, showing medium to strong correlations. Table 2 contains a small sample from our dataset and shows the ten most frequent verbs in the 'V *about* N' construction in the two study corpora. We see that four verbs (set in italics) are shared between top-10 lists, including the most frequent choices 'think' and 'talk'.

However, we also found differences between learner and L1 speaker verb choices, which can be explained on the basis of cross-linguistic transfer from Turkish, the learners' L1. To study cross-linguistic transfer effects, we worked with the 50-million word Turkish National Corpus (TNC) and looked up translation equivalents of VAC components (selected verbs and prepositions) in this corpus of L1 Turkish writing. Learners show both morphological and collocational transfer from their L1. For example, two verbs that occurred repeatedly in the 'V *about* N' construction in the Turkish learner data but not in the BNC were 'mention' and 'discuss' ('mention about sth.' and 'discuss about sth.').[2] We found that these phrases might be modelled on a construction in Turkish that consists of a noun with a postposition plus a verb ('N + postpos V') as in 'N *hakkında tartış-*', which translates to 'N about discuss'. In our ICLE Turkish data, we also observed a comparatively high frequency of 'V *in* N', the most common of the 19 analyzed VACs in terms of overall token number. This 'overuse' of 'V *in* N' (compared with L1 usage), could be due to the common use of 'in' translation equivalents as morphological markers in Turkish (case markers, adpositions). In Turkish, translation equivalents of English 'in' often combine with nouns, as in '*şüphe içinde*' ('hesitation in (locative) inside of'), which may have led to 'be in hesitation' in learner production. Similar unidiomatic uses of this VAC in the Turkish learner data include 'parents are in lack of money' and 'they were in desire for big amounts of money'. So, what do these findings mean for teaching English as a second (L2) or foreign language to L1 Turkish learners? Based on our study, we think that, when we introduce constructions, it would help to highlight those verb-VAC combinations that are typical in usage. We would also recommend addressing cross-linguistic transfer effects by pointing out typological

[2]It should be noted here that instances of 'discuss about' have also been observed in spoken English as a Lingua Franca (Mauranen, 2012) and in varieties of English, such as Indian English (Mukherjee, 2009, p. 123) and Australian English (as pointed out by a reviewer of an earlier version of this paper).

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

differences between the learners' L1 and their L2, as well as to focus on constructions that learners struggle with. In the case of L1 Turkish learners, 'V *in* N' would be one of those constructions.

The second study in this section on usage-based SLA was co-authored by Jamie Garner and appeared as Römer and Garner (2019). The study focuses on the use of five high-frequency VACs (a subset of the 19 VACs covered in Römer & Yilmaz, 2019) in L1 Italian and L1 Spanish learner English production. Different from the previous study, we here focus on spoken rather than written learner data, and on data from different proficiency levels rather than just advanced learner data. Our aim here was to find out how learner VAC knowledge develops from low-intermediate (B1) to high-advanced language proficiency (C1/C2). We used a subset of the Trinity Lancaster Corpus (TLC; Gablasova et al., 2019), which is a cross-sectional corpus of English as a Foreign Language (EFL) test taker speech, as well as the BNC as a proxy for L1 English usage. Our TLC subset contained transcripts of speech samples from 750 learners, making up about 1 million words. From these corpora we extracted data on 'V *about* N', 'V *for* N', 'V *in* N', 'V *like* N', and 'V *with* N', and performed various types of analyses on the VAC data, including normalized entropy analysis, correlation analysis, and regression analysis. We found that learners' verb-in-VAC production becomes more predictable and more Zipfian[3] as their proficiency increases, bringing it more in line with target usage. Looking at three groups of learners (Italian, Spanish, Mexican) separately, we also found that these groups strongly overlap in their verb selections for the five focus VACs if proficiency levels are controlled for. Verb-in-VAC frequency in usage was found to be a strong predictor for learners' verb choices, and we observed an increase of overlap of verb choices with target usage as learner proficiency increased. This study again provides additional evidence for strong usage effects on SLA, which confirms findings of earlier, smaller scale studies on VACs (e.g., Ellis & Ferreira-Junior, 2009; Ellis et al., 2014). I would argue that we could use this knowledge in pedagogical contexts. Knowing that learners are sensitive to input frequencies could be taken advantage of in teaching, for instance by creating what Madlener (2015) calls 'structured input floods', which have been shown to enhance construction learning. In such input floods, learner input is enhanced (or flooded) 'with task-essential exemplars of the target structure in a principled way' (Madlener, 2015, p. 92).

The third and final study in this thematic section, Römer (2019), focuses more comprehensively on VACs than the two studies I just discussed. Rather than looking at a subset of constructions, I am trying to explore all VACs used by L1 German learners of English from beginner to advanced levels, using a syntactically parsed corpus (a corpus which contains information on syntactic relationships between words). The aims of this study are to find out what the first VACs are that emerge in these learners, what changes there are over time in the learners' VAC repertoires and in their preferred verb-VAC associations, and what role formulaic sequences play in their VAC acquisition. The study combines methods from both CL and NLP (Natural Language Processing) and uses data from a pseudo-longitudinal corpus of English EFL learner writing called EFCAMDAT, the Education First-Cambridge Open Language Database (Alexopoulou et al., 2015; Geertzen et al., 2013). The study also uses reference data from a subset of COCA of roughly 80 million words of academic writing.

What I study in the project that Römer (2019) is based on are the patterns of a large number of verbs (about 700,000 verb tokens), which were extracted from about 68,000 EFCAMDAT learner texts (roughly 6 million words) from CEFR levels A1 through C1. The project combines various types of analytic techniques, including comparisons of frequency-sorted VAC and verb-in-VAC lists across learner levels, correlation analysis, normalized entropy analysis, verb cluster analysis, and concordance analysis (details on these analytic techniques are provided in Römer, 2019).[4] The first sets of results from this large-scale study (summarized in Römer, 2019) indicate that the first VACs that

[3]Zipfian verb-in-VAC distributions follow Zipf's Law, which states that the frequency of words decreases as a power function of their ranks in the frequency table (Zipf, 1935). For VACs, this means that very few verb types account for the vast majority of tokens while most others occur only once or twice.

[4]I am grateful to a number of colleagues, all former students of mine, who supported me in carrying out these analyses. These colleagues include Kristopher Kyle, who I was able to win as an NLP consultant for the project and who worked with me on the compilation of VAC databases for each EFCAMDAT subcorpus; Jamie Garner, Olivia Higgins, and Kristina

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

emerge in L1 German learners of English include simple transitive, intransitive, and copula *be* constructions, that the learners' VAC repertoire grows steadily from A1 to C1 levels, and that from low to higher proficiency levels, verb-in-VAC preferences shift from predominantly high frequency verbs to lower frequency verbs. Results for two high-frequency VACs that I have studied in more detail, the transitive construction (subject + verb + direct object) and a simple modal verb construction (subject + modal verb + verb + direct object), also show that from lower to higher levels, verb associations move closer towards an L1 usage norm, showing more lexical overlap with and more similar entropy values to the COCA data.

An additional finding from the verb cluster analysis that I found particularly interesting as a phraseologist was that we can observe a development from small sets of fixed word sequences at lower proficiency levels to a wider repertoire of more variable sequences at the B1 and B2 levels. So, we see formulaic sequences giving way to more abstract constructions as learners advance in language proficiency. For example, while A2 level learners repeatedly produce 'make me happy', B2 learners use variations of this phrase such as 'make you happy', 'would make you happy', and 'will make you happy' in addition to 'make me happy'. Similarly, upper-intermediate and advanced learners who contributed to EFCAMDAT do not only frequently use 'make a decision', but also repeatedly produce 'make difficult decisions', 'make the right decision', and 'make certain decisions'. This indicates a shift from fixed phrases to more variable constructions.

3.2. Language assessment

I will now move on to phraseology and corpus studies within the area of language assessment and feature two studies that were published as Römer (2017) and Römer and Banerjee (2018). The first of the two studies focuses on n-grams and p-frames in spoken English (i.e., items that lie at the intersection of lexis and grammar), and also examines rating scales of high stakes speaking tests. Its aims are to use phraseological analysis to illuminate aspects of speech, and to determine in how far rubrics of high stakes speaking tests capture these aspects of speech. For this study, I used two corpora of spoken English: MICASE, the Michigan Corpus of Academic Spoken English, and the spoken subcomponent of the BNC. I applied various corpus-analytic techniques to these corpora, with the goal of identifying central aspects of authentic spoken lexicogrammar. The analyses highlighted core features of speech, such as the frequent use of discourse markers (e.g., 'so', 'like'), backchanneling devices (e.g., 'yeah', 'okay'), and hesitation markers (e.g., 'uh', 'erm'), and also showed that spoken language is dominated by phraseological patterns which perform central discourse functions and connect lexis and grammar. To illustrate this, Table 3 shows the 20 most frequent fixed three-word sequences (3-grams) in MICASE and the spoken part of the BNC (BNC_spoken). If we try to categorize those 3-grams and see what they serve to do, we find that they function variably as quantifiers (e.g., 'one of the', 'a bit of'), discourse markers (e.g., 'you know', 'I mean'), discourse structuring devices (e.g., 'in terms of', 'at the same time'), and stance markers (e.g., 'I don't think', 'I think that').

In my analysis of rating scales of the speaking components of internationally recognized tests (e.g., TOEFL iBT, IELTS, Cambridge English: Advanced), I found that these aspects of authentic spoken English were insufficiently reflected in these rating scales, and that there is a lack of acknowledgement of the interrelatedness of vocabulary and grammar there. My recommendation at that point, purely driven by corpus evidence, was to avoid giving separate scores for two skills (i.e., lexical and grammatical competence) that are closely interrelated. This recommendation is in line with Alderson and Kremmel (2013), which questions the usefulness of separating vocabulary and grammar in the context of testing learners' reading ability. I argued that a phraseological approach could help bring speaking assessment more in line with the reality of authentic speech, and also help make life easier for raters. On this last point, a study by Ruegg et al. (2011) showed that raters of a proficiency test in Japan, the

Sherrill, who helped me with the analysis of verb clusters of selected high-frequency verbs; and Aidan Coleman, Riley Litzenberger, and Arly Salinas, who assisted me in preparing VAC data for the correlation and entropy analyses.

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

Table 3. Top-20 3-grams in MICASE and BNC_spoken (table adapted from Römer, 2017, p. 483)

	MICASE		BNC_spoken	
Rank	3-gram	Frequency	3-gram	Frequency
1	a lot of	1,199	a lot of	4,513
2	I don't know	1,107	I don't know	3,932
3	one of the	754	I don't think	2,943
4	you have to	610	one of the	2,549
5	this is the	566	do you want	2,013
6	a little bit	542	what do you	1,910
7	this is a	498	and I think	1,813
8	in terms of	480	you want to	1,745
9	I don't think	473	be able to	1,656
10	I think that	410	the end of	1,618
11	be able to	391	you have to	1,609
12	and this is	390	it was a	1,599
13	you have a	377	do you think	1,544
14	you know what	377	I think it's	1,538
15	I mean I	363	a bit of	1,537
16	part of the	338	going to be	1,535
17	some of the	331	I think it	1,524
18	you can see	326	I mean I	1,491
19	you need to	326	I think that	1,377
20	you know the	321	there was a	1,360

Kanda English Proficiency Test, found it difficult to keep lexis and grammar apart in scoring tests and often gave the same score for both areas (see also Ruegg, 2015). In the conclusion to Römer (2017), I noted that there is a need to make concrete suggestions on how to implement a phraseological approach in rating scale development, and that, as a corpus linguist, I did not have the relevant expertise to make those suggestions but would need to team up with an assessment expert.

So, for the next featured study, Römer and Banerjee (2018), I did team up with a language assessment professional, Jayanti Banerjee. This study focuses on frequent n-grams and p-frames in spoken test taker language across proficiency levels, and aims to validate the MET (Michigan English Test) speaking test with the help of phraseological analysis. In this study, we asked whether, and if so in what ways, test taker phraseological competence differs across levels, and whether the MET speaking test provides a good way of eliciting what learners are able to do at each level in terms of using phraseological items. To address these questions, we created a 448,000-word corpus of 720 transcribed MET speaking tests from four CEFR levels (A2, B1, B2, C1) and extracted lists of n-grams and p-frames of different lengths from the level-specific subsets of this spoken test taker corpus. Our findings indicate: (i) an increase of repeatedly used n-gram types from levels A2 to C1, (ii) a development from simple, and at times unidiomatic sequences (e.g., 'to make exercise', 'it's more easy'), to more complex, idiomatic phrases, and (iii) strong differences in the types of high frequency n-grams used across levels. For instance, at the A2 level we find n-grams that include a number of hesitation markers ('erm I think', 'erm there is a') and repetitions ('you can you can', 'in the in the'), whereas we

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

find several high-frequency n-grams expressing stance or evaluation at the B2 and C1 levels ('I think that', 'it seems to be', 'on the other hand').

Another interesting thing we observed in this study, in terms of the development of test taker phraseological competence, was an increase in the productivity of p-frames at the higher levels. While learners at the B2 and C1 levels used p-frames such as 'I * that' productively with a high number of different lexical items filling the '*' slot (including 'think', 'know', 'believe', 'remember', 'understand'), A2 level learners mostly relied on a small number of p-frame realizations, with one of them being the dominant choice (here 'I think that') and accounting for the lion's share of all p-frame tokens—illustrating what Ellis (2012, p. 17) calls a 'phrasal teddy bear'. Overall, our findings suggested that the MET rating scale is successful at distinguishing between test takers at different levels of phraseological competence, and that the MET speaking test design enables test takers to demonstrate their language proficiency insofar as this is evidenced by increases in their phraseological competence.

3.3. Empirical legal language analysis

The last two studies I would like to share with you fall into the 'beyond' category of my presentation title and provide examples of empirical legal language analysis. So, how can corpora and corpus approaches (again with a focus on phraseology) be used in legal scholarship to help us better understand problems that are of concern to lawyers and legal scholars?

The first of the two studies in this section, published as Ren et al. (2020), is the result of collaborative work between a group of lawyers and linguists, and focused on a phraseological analysis of a corpus of founding era American English. The aim of the study was to determine the original public meaning of the word 'case' in Article III of the US Constitution. We wanted to find out, with the help of corpus analysis, whether the meaning of 'case' during the founding era was limited to 'injured plaintiff litigation' or whether it referred to something more general. This former, more narrow interpretation was used in 2019 by the US Court of Appeals for the Fourth Circuit to justify dismissing a lawsuit brought against Donald Trump claiming that he violated the Constitution's prohibition against receiving emoluments from foreign states. To address this question, we used data on 'case' (and 'cases') and the patterns the word frequently occurs in retrieved from a corpus called COFEA, the Corpus of Founding Era American English, consisting of texts from 1760 to 1799 and making up more than 136 million words. Our COFEA analysis was complemented by historical text analysis for which we accessed the Founders Online database (founders.archives.gov), a text repository that stores texts from 1706 to 1836.

Results from this study, based on an analysis of the phraseological patterns in which 'case' typically occurred in founding era American English, indicate that those who drafted and ratified the US Constitution likely would have understood 'cases arising under laws passed by the general Legislature' to be a type or an example of 'questions as involve the National peace and harmony'. In other words, the latter ('questions…') would have been a more general category of jurisdiction than the former ('cases … '; Ren et al., 2020, p. 557). The results also indicated that those who drafted and ratified the US Constitution would not have understood 'case(s)' as having a stable, inherent meaning such as 'adversarial litigation', but understood that the word takes on different meanings, constructed through its combination with accompanying words. We argued that 'case' functions as a so-called 'shell noun' (Schmid, 2000), a semantically general or vague noun that inherits meanings from the specific lexical contexts it is used in. This work in applied CL is potentially important in terms of its practical relevance, as it may prompt a reevaluation of the Supreme Court's assumption that the original meaning of 'cases' in Article III was restricted to meaning 'injured plaintiff litigation', an assumption that we cannot confirm based on our extensive corpus analysis.

Continuing the collaboration between a lawyer (Clark Cunningham) and a corpus linguist (myself), the second study in this section—and the last one I would like to showcase in this paper—again focuses on a phraseological analysis of a corpus of founding era American English, this time with the aim to determine the original public meaning of the word 'misdemeanor' in Article II, Section 4 of the US Constitution. We chose this topic because the meaning of the phrase 'high crimes and

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

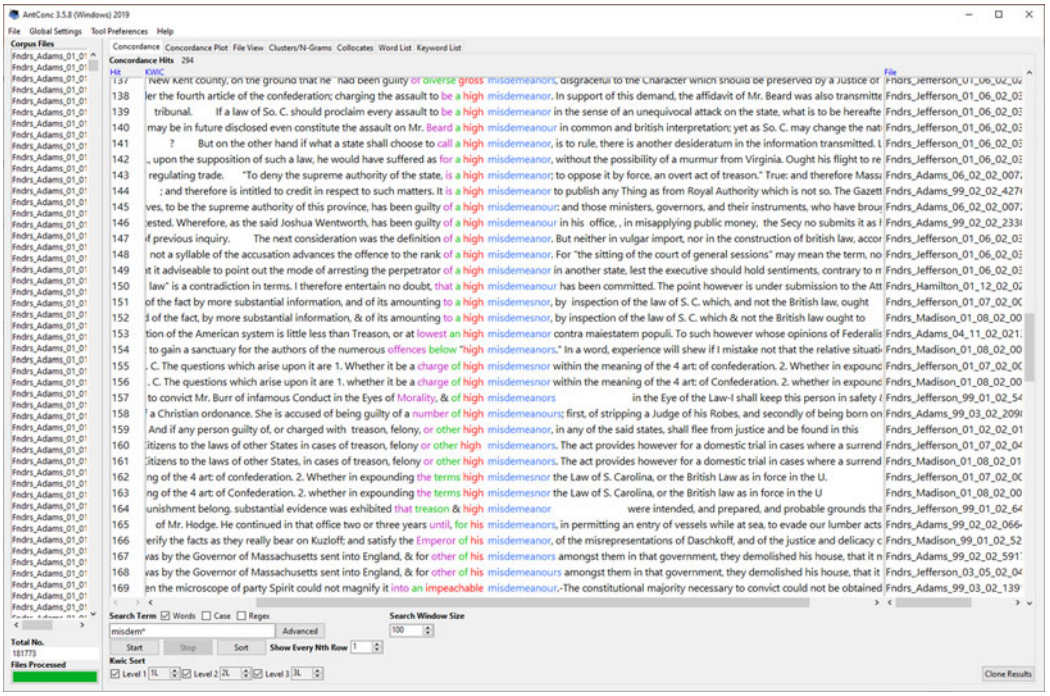


Figure 2. Part of the concordance results for the search term 'misdem*' in the Founders Online corpus, sorted to the left.

misdemeanours' in the section of the US Constitution that defines grounds for impeachment has been a fairly hotly contested issue of public discourse in this country over the last two years, and an issue that legal scholars do not appear to agree on. For instance, emeritus Harvard Law professor Alan Dershowitz argued in January 2020 during the first Senate impeachment trial of Donald Trump that the terms 'crime' and 'misdemeanor' are synonymous and that what Trump was accused of did not constitute a crime. Several legal scholars rejected Dershowitz's interpretation and argued that the phrase 'high crimes and misdemeanors' was a legal term of art or idiom. Our study was first presented at the 6th Annual Law and Corpus Linguistics Conference at Brigham Young University in February 2021 (Cunningham & Römer, 2021) and is currently in preparation for publication in a law journal.

For this study, we extracted data on 'misdemeanor(s)' and the patterns it frequently occurs in from a 67 million word corpus of more than 180,000 texts downloaded from the Founders Online archive. The texts in our corpus were written between 1706 and 1836 by or to early American leaders, including George Washington, Benjamin Franklin, and Alexander Hamilton. We used this corpus to address several open questions in legal scholarship on this topic, including: (i) In founding era American English, was 'high crimes and misdemeanors' a fixed phrase or legal term of art? (ii) Were 'crime' and 'misdemeanor' used synonymously during this time? (iii) Did 'high' and 'other' in sequences in which they are followed by two nouns modify both nouns or only the first one? and (iv) What meaning does 'high misdemeanors' add to the impeachment clause? A search for 'misdem*' in *AntConc*, using the * character to capture spelling variants as well as singular and plural forms, resulted in 294 instances of the search term in the Founders Online corpus, distributed fairly evenly across 203 texts. A subset of these 294 results is displayed in the *AntConc* screenshot in Figure 2. Sorting the context of the concordance allowed us to easily see phraseological patterns around the search term. To complement the concordance search, we carried out a distribution analysis to find out how 'misdem*' is distributed across texts, an n-gram analysis to extract repeated phrases of

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

different lengths that contain 'misdem*', and a collocation analysis to identify which words are most strongly associated with 'crime(s)' and 'misdemeanor(s)' and contribute to their meaning in context.

The findings of these analyses indicate that in founding era American English usage, 'high crimes and misdemeanors' was not a fixed phrase or a legal term of art but that there was considerable variability in this phrase (e.g., 'crimes or misdemeanors', 'a high misdemeanor', 'misdemeanors, and other crimes'). Our results on the collocation profiles of 'crime(s)' and 'misdemeanor(s)' suggest that the two nouns were not used synonymously but occur in distinct lexical contexts. We also found that, after we looked systematically at the construction 'other N and N' in our corpus, 'other' tends to modify both nouns, as in 'other officers and soldiers', 'other days and times', or 'other matters and things'. The same applies to 'high' in 'high N and N' sequences, in which it predominantly modifies not just the first but both nouns (e.g., 'high consideration and respect', 'high regard and esteem'). So, to those who ratified the US Constitution, the impeachment clause would have been interpreted to include 'treason, bribery, and other high crimes and *other high* misdemeanors'. Our analysis of what meaning 'high misdemeanors' adds to the impeachment clause is still ongoing as it requires us to go back and forth between the concordance results and the 203 source texts in which 'misdemeanor' occurs. Initial results from this ongoing analysis suggest that 'high misdemeanor(s)' repeatedly appears in contexts describing misconduct that affects government functioning or threatens the authority or security of the state.

4. Corpora for AL: Looking ahead

As I am coming to a close, I hope that this journey through some of my recent work has shown that corpora and corpus analytic techniques can be powerful tools in addressing important questions across a range of different areas of AL (and beyond). My hope was to illustrate how corpus-based research with a focus on phraseology has led me to new insights into the process of language acquisition, and the patterned nature of various types of language, and that those insights have implications for SLA, for pedagogy, for language assessment, and potentially for other fields of research (such as legal scholarship) too. Another thing I hope to have highlighted in this paper is the importance of inter- and intra-disciplinary collaboration that has significantly impacted and strengthened my recent research. I would argue that collaborative work is essential for an impactful future of AL. My research in applied CL has benefited immensely from collaboration and from the additional skill sets that each collaborator brought to the table. In my recent research history, this included the skill sets of language acquisition researchers, language teachers, language assessment specialists, and legal scholars, including Jayanti Banerjee, Clark Cunningham, Nick Ellis, Jamie Garner, and Selahattin Yilmaz, all of whom I am lucky to count among my co-authors.

Looking ahead at next tasks that would be important to include in our applied CL research agenda, I would recommend that we work more on creating and analyzing new types of corpora including longitudinal learner corpora, dense learner corpora (corpora that contain a large amount of data from each learner), and English as a Lingua Franca (ELF) corpora, following the lead of the ELFA (English as a Lingua Franca in Academic Settings), VOICE (Vienna Oxford International Corpus of English), and ACE (Asian Corpus of English) projects. I would also suggest that, in future research in this area, we continue to combine different types of data and analytic methods. The studies I showcased here all benefited from a combination of quantitative and qualitative research methods. In other studies, I have argued in favor of triangulating corpus evidence and data collected in psycholinguistic experiments, which can provide a different lens on the phenomenon of interest and strengthen the validity of a study (Römer et al., 2014, 2018). A third recommendation would be to team up with researchers from neighboring fields (e.g., psychology, computer science) and to connect more to other disciplines that may benefit from empirical language analysis and approaches to highlighting patterns in language. In my recent interdisciplinary collaborations, I have learned that language patterns can provide answers to important questions unrelated to linguistics, and also that the 'language' of interest may not always consist of words. In a project that aimed to uncover central aspects of jazz improvisation, my collaborator, music professor Martin Norgaard, and I worked with a corpus of jazz

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

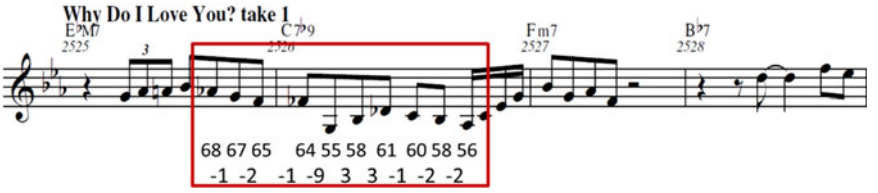


Figure 3. Example of an interval sequence from Charlie Parker's 'Why do I love you?'.

solos in which pitches were converted into intervals represented as positive or negative numbers of half step changes (e.g. −1, 3), as illustrated by the example from Charlie Parker's 'Why Do I Love You?' in Figure 3. With the help of corpus analysis (using *AntConc* on a corpus of 456 solos from the Weimar Jazz Database), we demonstrated the ubiquitous use of patterns by jazz artists and highlighted the central role patterns play in real-time music creation (Norgaard & Römer, forthcoming). I would encourage fellow applied linguists to explore potential opportunities for interdisciplinary collaboration and see how their skills can be applied to address questions in other research domains.

References

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556.

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, *1*(1), 96–129.

Anthony, L. (2019). *AntConc (Version 3.5.8) [Computer Software]*. Waseda University. http://www.laurenceanthony.net/software.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Cunningham, C. D., & Römer, U. (2021). Investigating the original public meaning of 'misdemeanors' in the impeachment clause. *Paper presented at the 6th Law and corpus linguistics conference*, 5 February 2021. Brigham Young University, Provo, UT.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 188–221. https://doi.org/10.1075/arcl.7.08ell

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second language processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*(4), 405–431. https://doi.org/10.1075/lab.4.4.01ell.

Fletcher, W. H. (2007). *KfNgram*. USNA. Retrieved from http://www.kwicfinder.com/.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–158.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). *Proceedings of the 31st second language research forum (SLRF)*. Carnegie Mellon University: Cascadilla Press.

Goldberg, A. E. (1995). *Constructions. A construction grammar approach to argument structure*. University of Chicago Press.

Hunston, S. E., & Francis, G. (1998). Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics*, *19*(1), 45–72.

Hunston, S. E., & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Madlener, K. (2015). *Frequency effects in instructed second language acquisition*. De Gruyter Mouton.

Mauranen, A. (2012). *Exploring ELF. Academic English shaped by non-native speakers*. Cambridge University Press.

Mukherjee, J. (2009). The lexicogrammar of present-day Indian English: Corpus-based perspectives on structural nativisation. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar interface* (pp. 117–135). John Benjamins.

Norgaard, M., & Römer, U. (Forthcoming). Patterns in music: How linguistic corpus analysis tools can be used to illuminate central aspects of jazz improvisation. *Jazz Education in Research and Practice*.

Ren, H., Wood, M., Cunningham, C. D., Abbady, N., Römer, U., Kuhn, H., & Egbert, J. (2020). 'Questions involving national peace and harmony' or 'injured plaintiff litigation'? The original meaning of 'cases' in Article III of the Constitution. *Georgia State University Law Review*, *36*(5), 535–605.

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press

Römer, U. (2008). Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing. *Functions of Language*, *15*(1), 115–130.
Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140–162.
Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477–492.
Römer, U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, *24*(3), 268–290.
Römer, U., & Banerjee, J. (2018). Validating the MET speaking test through phraseological analysis: A corpus approach to language assessment. *CaMLA Working Papers*.
Römer, U., & Garner, J. R. (2019). The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency. *International Journal of Learner Corpus Research*, *5*(2), 207–230.
Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, *38*, 59–79. https://doi.org/10.2478/icame-2014-0006.
Römer, U., Skalicky, S., & Ellis, N. C. (2018). Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks'. *Corpus Linguistics and Linguistic Theory*. Advanced online access. https://doi.org/10.1515/cllt-2016-0055.
Römer, U., & Yilmaz, S. (2019). Effects of L2 usage and L1 transfer on Turkish learners' production of English verb-argument constructions. *Vigo International Journal of Applied Linguistics*, *16*, 107–133.
Ruegg, R. (2015). An experiment in the ability of raters to evaluate lexis in writing. *Language in Focus Journal*, *1*(1), 38–50.
Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*(1), 63–80.
Schmid, H.-J. (2000). *English abstract nouns as conceptual shells. From corpus to cognition*. De Gruyter Mouton.
Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. McH. (2004). *Trust the text. Language, corpus and discourse*. Routledge.
Sinclair, J. McH. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 407–410). John Benjamins.
Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics*, *30*(1), 115–137.
Zipf, G. K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. The MIT Press.

Ute Römer is an Associate Professor in the Department of Applied Linguistics and ESL at Georgia State University. Her research interests include phraseology, second language acquisition, academic discourse analysis, and the application of corpora in language learning and teaching. She serves on a range of editorial boards of professional journals and is General Editor of the book series *Studies in Corpus Linguistics* (John Benjamins). Her research has been published in a range of applied, corpus, and cognitive linguistics journals including *Language Learning*, *The Modern Language Journal*, *International Journal of Corpus Linguistics*, *Corpora*, and *Cognitive Linguistics*. More information about her work can be found at: http://uteroemer.weebly.com/.

Cite this article: Römer, U. (2022). Applied corpus linguistics for language acquisition, pedagogy, and beyond. *Language Teaching*, *55*(2), 233–244. https://doi.org/10.1017/S0261444821000392

https://doi.org/10.1017/S0261444821000392 Published online by Cambridge University Press