Teacher-Expert Dialogues

Opportunities and Synergies for Structured Corpora, AI and TESOL: An Interview with Professor Mark Davies

Stephen Jeaco
Xi'an Jiaotong-Liverpool University, China
(Email: steve.jeaco@xjtlu.edu.cn)

Received: 2 December 2025; Accepted: 19 March 2026; Published: 5 April 2026
https://doi.org/10.58304/tc.260302

Abstract
Professor Mark Davies created the Corpus of Contemporary American English (COCA), making a huge contribution to English language description and teaching, providing fresh insights into differences between British and American English, changes over time, and contrasts across genres. He also developed a corpus architecture and corpus tool interface that provides easy yet powerful access to COCA and a host of other structured corpora. In an interview recorded a few weeks after the addition of Artificial Intelligence features across the English-Corpora.org platform, Professor Davies discusses the development of his corpora and his interface over the last two decades. He provides examples of how corpora have been and continue to be extremely useful for English language learning and teaching. The interview also includes some discussion of how and why the AI functions were developed, touching on some key considerations for prompt design and one's general approach to language data. The interview concludes with a summary of insights and recommendations, including a call to keep data front-and-center and to promote active learning through AI-assisted discovery.

Keywords
Corpus design, artificial intelligence, AI assisted corpus analysis, supporting second language learners

Introduction

Professor Mark Davies, now retired, was a professor of linguistics at Brigham Young University. He is well-known for creating structured corpora, his innovative work on corpus architecture and his powerful yet easy-to-use interfaces for corpus tools[i]. His most famous corpus is the Corpus of Contemporary American English (COCA) – a widely used resource for exploring change over time and variation across genres, with over 5,000 citations[ii] as of late 2025. He has also created corpora for (among others) historical American English (COHA), dialectal variation (GloWbe), systematically selected web pages (iWeb), up-to-date and ever-developing news (NOW) and for Spanish[iii] and Portuguese[iv]. He is a highly cited scholar with books, book chapters and journal articles on American English, Spanish, Portuguese, corpus design, historical change and phrasal analysis. In this interview, the focus is on language learning and teaching, especially the contribution Professor Davies has made to Data-Driven Learning (DDL) – the use of activities that present multiple, authentic examples to language learners in order to provide the raw data they need to formulate their own understandings of the patterns and usage of words, phrases and structures. For an overview of DDL and how it relates to the importance of educational philosophies, guided noticing and learner autonomy in language teaching, see Flowerdew (2015) or Pérez-Paredes and Boulton (2025). In the

interview, Professor Davies describes how his awe-inspiring, technological developments in corpus linguistics were steered by attention to their practical application for the language learning classroom – both in the past and in the new era of AI.

When did you first realise the huge potential of COCA for English language teaching?
It was very gradual. I'll just give you a little bit of background information that will let you understand how I really had no idea where all this would go. My doctorate is in Spanish and Portuguese - medieval Spanish and Portuguese syntax, if you can believe that! So, I was looking at syntactic change in Spanish 800 years ago. And as you know, there's not a lot of 800-year-old speakers, and so I had to just bring together texts and create a corpus to be able to look at change. But then I got interested in the field of corpus linguistics – I didn't even know there was such a field; there was methodology and so on. And in 2003, I moved from Illinois State University, where I had been teaching Spanish, to BYU, where I was teaching just general linguistics. And that's when I started creating corpora in English. I have no background in TESOL; in other words, I've never taught English to non-native speakers. I've taught English grammar and other courses in English linguistics at BYU many, many times, but it was to native speakers. So, honestly, I have no training in TESOL itself and it has just been little bits here and there that I picked up mainly by working with other people. So, it was kind of a strange roundabout way of getting into tools that are used in TESOL.

This is really tied into what I'm doing in English. Even though I had a doctorate in Spanish, there were linguistic issues that I found incredibly troublesome, especially with collocates and collocations – just getting the right word for a particular context. And so, having a corpus that could in two or three seconds give you the insights of a native speaker as a non-native speaker of Spanish and certainly Portuguese was invaluable. But that experience of interacting with data in a language that is not my own helps me have a lot more empathy for people learning English. Unfortunately, it's far too common that corpus linguists of English don't speak another language. They have little, if any empathy for what those learners and teachers are going through. Fortunately, I have some of that empathy because of Spanish and Portuguese.

So, in 2004, just as kind of proof of concept, because I had already created Spanish corpora, I had access to an open source version of the British National Corpus[v] from the BNC people. I purchased it and I thought, well, I'll just see if I can put it in my architecture and interface. And so for the first four years with the BNC – between 2004 and 2008 – I suppose the corpus was used for teaching and learning, but, honestly, I designed it as a tool for linguists. And even when I created COCA in 2008, I could only see things through the lens of a linguist; that's what I was. But I had a very good friend at BYU named Dee Gardner. And he did have a background in TESOL – he had taught English to non-native speakers – he was very interested in pedagogy, but he was also very interested in corpora. And so Dee and I started this partnership in about 2004, where I'd create tools and Dee would say, "Erm. Yeah, that's nice. But how's anyone ever actually going to use this? What difference would it make for teachers or learners?"

And so that collaboration was crucial because it helped me see things through the eyes of someone who was very practical about how are people actually going to use this. But the recognition that it could or would or was being used by language learners, that was just very gradual. If you had asked me in 2008, I would just have said, "Sure, I guess people can." By 2013, 2014, 2015, I suppose I understood it much better. But it was all very gradual.

But then for the last six or seven years, virtually everything that I've done with the corpora – virtually everything new that I've added – has been to make the corpora more useful for

language learners and teachers. Of course linguists use it – that's fine. But my emphasis has definitely been on creating tools for teachers and learners.

And so, looking over the last two decades, what do you think are the biggest changes in the ways that your platform has been used?
When I originally created COCA I was trained as a historical linguist, so I thought it would be interesting to see how things have changed since 1990 to very recent times. But more than anything with COCA I was looking at genres, so I wanted it to be something that would be comparable to the BNC. The BNC would be for British English, and this would be for American English. But in the back of my mind, as a historical linguist, I had been thinking, "I want to create a historical corpus." And I created a couple of stop-gap ones, but then in 2010, I created COHA.

And then, I thought, what's the one main thing that I could investigate or that I could create a corpus for relatively easily that would complement the genres and the historical. And, for me, at least, the answer was obvious – dialects. And so, the GloWbE was created in 2013. And so now I have genres, historical, and dialects. And everything that I've done since then has been variations on those three.

Perhaps one of the weak points of COCA compared to the BNC was the spoken part. It was never as informal as I wanted it to be. And so that's why in 2018-2019, I created the TV and Movies Corpora, which is extremely informal language. On many measures, it actually is more informal than data from actual spoken corpora, as paradoxical as that might seem. And the NOW corpus was to look a very recent change. So my main focus has been genres, historical and dialects, but I've created several others that have fed off of them.

Over the years, what are some of the most interesting uses that you've seen of corpus tools in terms of helping teachers understand language – the things in a workshop that seem to make the eyes of the whole audience light up?
I've done many workshops in many different countries, and it is always very satisfying to see people's eyes kind of light up. For me, two uses stand out. Firstly, there's exploring semantically; I have really tried to create corpora that are semantically searchable. You can actually include this into the search: search by synonyms or search by customized wordlists. And then the data that's displayed should provide really good insight into the meaning and usage of words and phrases. So that's always been front and center. Secondly, there's exploring grammar. Many people still look at language from a very prescriptive point of view, and this is particularly the case in Asia for cultural reasons. There they might say, "We want to speak and write the way they do in the UK or the way they do in the US, so let's get things right grammatically." But, grammar is often a matter of 50 shades of gray, where you go from ungrammatical to completely grammatical and all these gradations in between that really help you sound more native-like. And corpora can provide that really nicely – that insight into nuance in grammar. But obviously, word meaning is probably the most important. So, I'd say those are the two main things: semantic insight in terms of the meaning and usage of words or phrases, and a realistic, accurate, descriptive explanation of English grammar – not just what the textbooks say, but what actually is going on.

Maybe I could give you one example; I always give this example, but for me, it's a very nice example of what corpora can offer. When I first went to China, I noticed that people would use the word *seldom* a lot. "We seldom do that." "I seldom have time." And people's English in

China was just amazingly good – much better than I initially expected. But there was this one word – *seldom* – that was used in a way that just seemed awkward to me.

So, in a corpus, very easily in COCA, you can do one search that tells you *seldom* is used much more in formal speech than informal speech – so genres are important. You can also see with the same search that *seldom* is really decreasing in usage in American English over time. And then if you go to another corpus, you can see very easily that *seldom* is used much more in British English than in American English. And so just with a couple of searches, you can see if I use *seldom*, I'm going to sound old-fashioned, formal, and British! There's nothing wrong with any of those things; it's just if you're trying to sound colloquial in American English, it's the wrong word to use. No matter how good a dictionary is, it is not going to provide you with that level of nuance and granularity: genres, formality, historical change, dialectal variation. But four or five seconds with corpora and you have all of that. There's just no way you can get that from a dictionary; a corpus is invaluable for that kind of thing.

In the last two years you've contributed greatly to the discussion on how corpora should still be relevant in the new age of Artificial Intelligence. Can you summarize the main thoughts from your papers and videos - specifically, those most relevant for language learning and teaching?

There's just a couple of general comments that I wanted to make on AI first. I think we all realize that AI has revolutionized other fields of inquiry. You're probably familiar with AlphaFold from DeepMind using AI, for which they actually won the Nobel Prize for chemistry of all things last year, despite the fact that they were using AI to look at protein folding. And you can look at many different fields where AI has just revolutionized things; AI is extremely good at looking at patterns. And the resources that I have online about corpora and AI contain a number of examples of this. Regarding collocates, for example, if you went to English-Corpora a year ago or Sketch Engine today, it will just give you a big, long list of collocates for a given word. And if you're really lucky, it will be grouped by part of speech: noun, verb, and so on. For a language learner, they're still just looking a big, long list of words, but the goal has always been to provide a semantic characterization of what these collocates mean, in order to tell us what the node word means and how it is used. AI is just amazing at grouping those collocates semantically, for me, that's the most obvious use case for AI.

But I'm becoming increasingly convinced that maybe the most useful and the most powerful use of AI is for KWIC (Key Word in Context) displays. Typically, in the past, you'd search for a word or phrase, and you'd get all these lines of text showing that word or phrase in context, but honestly, that's just overwhelming for non-native speakers. There's all these words/phrases in those lines of text that they don't understand. These learners are not corpus linguists! They're not going to be able to really analyze these lines of text. But if you can feed 100 or 200 lines into AI, and have it analyze those lines of text in three or four seconds, it will give you amazing insight – into not only the meaning and usage of that word or phrase, but also the patterns in which it occurs: semantic prosody (is this positive/negative) and pragmatics, and much more as well. It analyzes it from nine or ten different angles. And what's really powerful is that it can do that in any one of 30 different languages. So, imagine you're a language learner from China; you look up a word or phrase, you've got all these lines of text – almost impossible to analyze. But then you have AI analyze it for you and explain it all in Chinese. That just really revolutionizes the corpora for the end user, I think.

So, we've just been talking about the new functions on your platform for AI assisted analysis. So, what were some of the pedagogical considerations that influenced the design of those?

One thing is I wanted for AI to not be overly intrusive. Everyone nowadays is saying "AI this"; "AI that"; "Try our AI." And honestly, it's just becoming a little bit bothersome – everyone talking about "AI insights." So, first of all, I wanted the corpus data to remain front and center, but, on almost any page that has corpus data, there's just a button that says something like, "analyze this with AI." And then that will open up in another tab, and so you have the corpus data on one hand, and the AI explanation or categorization in another tab, and you can go back and forth, back and forth between the data and the explanation. And for me, I never wanted the corpus data to get lost.

But the whole time I've been doing this for the last year, my overriding question has been, if I am a non-native speaker in China, Vietnam, or Saudi Arabia or wherever, what kinds of corpus data are hard for me to understand, but will suddenly be a lot more manageable with AI analysis and categorization? So, as I've created all of these different hooks with AI; it's kind of the spirit of Dee Gardner still. The question is, "Cool. That's really cool that AI can analyze things, but at the end of the day, how is it going to make a difference for non-native speakers?" And the Key Word in Context and the collocates, especially, and comparing collocates, different words via collocates, those for me, are just the most obvious cases of "Wow, if I were a non-native speaker, this is what would really make the corpus data useful for me!"

COCA and your contributions to Corpus linguistics are so widely known and discussed that the large language models we use today must 'know' something about them. If you could inject some knowledge or a different perspective directly into an LLM, what would you want it to 'know' that it doesn't seem to already 'understand'.

First, we're talking here about what large language models understand about corpora or about language. I spent around six months – from November 2004 to March 2025 – looking at what LLMs predict about language: putting words in order in terms of frequency or for a phrase like *dark* + noun or adjective + *industry*. What is surprising is LLMs give you answers that, on the surface, look really plausible. They're not hallucinating. It's not they just made that up from out of thin air. They really look like "the right answer." But then when you compare it with corpus data, often they're wildly off the mark. So, there is a real temptation for language learners to just say, "You know what, I'm not going to use structured corpora; I'll just ask Gemini or ChatGPT or Claude. What are the most common adjectives that occur with this noun?" or "What are the most common collocates of this word?" The LLM will give you an answer that will seem plausible, but chances are very good that it won't actually agree with what's going on in a corpus. And then you have this kind of epistemological question: well, do I go with what the LLM said, or do I go with corpus data?

So lots and lots of people are getting lost in that sense that they're just saying, "Forget about corpora - that's really old school; we'll just ask ChatGPT." And you know, it doesn't agree with corpus data very well.

But this kind of basic epistemological question of what is the future of corpora in a world that is increasingly dominated by AI is something that if it doesn't keep corpus linguists awake at night, it should, because I really fear that corpus linguistics as a field downplays the importance or the value of AI; like, "It has nothing to offer; we're doing fine; nothing to see here; move along!" But this is dead wrong. End users are using AI more and more and more, and we can't

just ignore it. That's why I decided to integrate it; take the very best of AI and integrate it with really good linguistic data.

And so, moving on to think about sort of would-be prompt engineers or people developing AI workflows with Data-Driven Learning in mind… I was just wondering if you could tell us about some of the challenges or roadblocks that you encountered when you were trying to do that integration – some of the stuff that didn't work, some of the stuff that ended up on the cutting floor, if you like.

Let me give you what I think is a really interesting example – at least it was for me. For the Key Word in Context display, originally the prompt was just "Find any interesting patterns that you see in terms of the usage of this word or phrase." And I was getting really poor responses. So eventually, I asked ChatGPT and Gemini and Claude, "What would be a prompt that you think would really highlight interesting patterns? What are seven or eight or nine different angles that we should take to look at this word or phrase in context?" And the results were amazing! It wasn't me. It was the AI itself that said, "Oh, you want really good insight into the patterns in which a word occurs? Okay, have the prompt ask for this and this and this." And it was the difference between night and day. All of a sudden, those Key Word in Context answers were amazing, even for me, as a native speaker and as a trained linguist. Almost every time I have AI analyze a word or phrase in Key Word in Context, I learn something. And if I'm learning something, as a native speaker, a trained linguist, I'm sure that non-native speakers are also learning a lot. One thing that I didn't mention is that it's also possible via the interface, besides getting the explanation in one of 30 different languages, to choose one of 14 different user categories: "I'm an intermediate language learner", "I'm a corpus linguist", "I'm a translator", whatever. And then the answer that it gives you will be targeted towards that level. So, imagine you're from China and you've told it, "I'm a learner and I want the answer in Chinese." And so, it's going to be targeted right to your level of understanding –still, this really great insight, but targeted directly to you. That is what makes it so powerful!

In terms of designing those prompts and working out some of the most fruitful ways to get useful summaries, was using multiple agents really important?

It was important. Anytime I have a really important question where I need the AI to get it right, whether it's talking about linguistics or philosophy or history or whatever, I nearly always ask all three. For me, the main LLMs are: ChatGPT, Gemini, and Claude from Anthropic.

Has token limits been a factor?

Yes. It's a really big deal. Ideally, with Key Word in Context I would send it a thousand concordance lines, but very quickly it becomes impossible financially to have that happen. And so, for both the tokens that I'm sending and the tokens in the response I need to consider limits. If it's explaining some data I will purposely state in the prompt "Limit your answer to, (let's say), 150 words."

There's another reason for that. If you're a language learner and you get a 2,000-word response, that's just unmanageable; you can deal with 150 words, especially if it's in your own language.

I recently gave a presentation to a group of corpus linguists based in the UK, and they were saying, "Well, it seems like sometimes the answers are too short." And I explained that they are on purpose; it's not going to write a paper for you about the use of the passive in contemporary American English, because it's limited to 150 words on purpose. So, I've thought a lot about the prompt and the number of tokens going and the number of tokens coming back.

Reflections and implications for TESOL

In this interview, Professor Davies has highlighted a number of important issues, and he has provided a positive perspective on how AI can facilitate even more powerful interactions with corpus data for the purposes of language learning and teaching. In contrast to non-engagement attitudes towards AI in some current discussions on corpus linguistics, his corpus tools have recently been updated to harness the power of AI in different ways, and COCA and the host of other corpora are consequently now more accessible to language learners and teachers than ever. This synergy is very different from approaches trying to remove direct engagement with data from the language learner's interactions; keeping data front-and-center is something we should learn from and reflect on in our own teaching practice. We should remember that telling students about patterns of usage is not the same as teaching them how to notice these patterns in data for themselves; English-corpora.org keeps a balance between scaffolding language learners as they try to engage directly with authentic language data, and ensuring that their eyes are directed back to patterns and concrete evidence. Professor Davies has not only provided many investigations into the similarities and differences between corpus evidence and AI generated 'evidence'; he has also provided a platform and an approach that allows others to explore new ways of promoting deep inductive learning in our learners through AI assisted discovery.

Notes

1. https://www.english-corpora.org/
2. Google Scholar, 26 November 2025.
3. https://www.corpusdelespanol.org/
4. https://www.corpusdoportugues.org/
5. http://www.natcorp.ox.ac.uk/

References

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15-36). John Benjamins.

Pérez-Paredes, P. & Boulton, A. (2025). *Data-driven learning in and out of the language classroom*. Cambridge University Press.

Stephen Jeaco, PhD, is an Associate Professor at Xi'an Jiaotong-Liverpool University, China. He has worked in China since 1999 in the fields of EAP, linguistics and TESOL. His research background includes full stack software development for language and linguistics, corpus design, data architecture, corpus methods and educational and assistive technologies.