entreviSta

TONY MCENERY

por Hadinei Ribeiro Batista (PhD student / CAPES / UFRJ / Lancaster University)

Tony McEnery is Distinguished Professor of English Language and Linguistics at Lancaster University and Director of ESCR (Centre for Corpus Approaches to Social Science), which was awarded a Queen's Anniversary Prize for its research in 2015.

Revista Linguíʃtica: **How do you see the relationship of the formal model of language with the methodology of Corpus Linguistics?**

Tony McEnery: For me there is a very natural marriage of the two. If you have a theory of language, it should account for what people say and do in response to people saying things to them. So, abstract theories of language which don't account for everyday speech and writing hold little interest for me. I appreciate that other features do impact upon our speech and our writing but I don't think that they impact so strongly that a theory of linguistics should not seek to account for what people say and write. For me the link between Corpus Linguistics and formal models of language is very strong because a corpus gives evidence for what people should be trying to model with their theory. So any theory that cannot account for what people view as well formed utterances is deficient in my opinion. You don't problematize language to make language fit your theory. You problematize theory to make it fit language. So Corpus Linguistics gives very good evidence for theories of language in terms of setting parameters for what they must account for.

Revista Linguíʃtica: **Which implications does Corpus Linguistics have to Sociolinguistics? What are the points you consider most critical of Traditional Sociolinguistics?**

Tony McEnery: Again, I think it is very helpful to look at corpus data if you have the right type of corpus. For a corpus to be relevant to sociolinguistics, you obviously have to model or mark up features within that

corpus that allow sociolinguistic investigation. So, if you have a corpus, say, within which the gender of the speakers or writers isn't known, the age of the speakers or writers isn't known, the social class of the speakers or writers isn't known, or the location geographically, or all sorts of information that is relevant to sociolinguistics is absent, it is of very little interest to sociolinguists. Yet if on the other hand you have a well composed corpus, well balanced, which is relevant to sociolinguistic study, and within it there are lots of sociolinguistic variables reliably encoded in the data, then, in principle that is very interesting to sociolinguistics. If you have a corpus representing a wide range of speakers like the spoken BNC (British National Corpus) and within that you have information that is relevant for sociolinguistic questions, then it is obviously a useful source of data for sociolinguists. Now, there are certain types of sociolinguistic studies that seem to be less well served by corpora in my view. So, people who are interested in phonetics for example. There is very little available to them from Corpus Linguistics, I think. So, it's only now, for example, 20 years later, that people are starting to get access to the recordings of the spoken BNC from 1994, which explains why I don't think I have seen it used once for sociophonetics research. Maybe some types of sociolinguistic studies are less well served than others. Sociophonetics, I think, is probably an example that. Though with that said, some sociophoneticians, like Paul Kerswill, have recorded and transcribed data under slightly limited circumstances or in laboratory conditions. They have built things that look like corpora, specifically designed for sociophonetic study. Sociophonetics is an interesting and developing area I think.

Revista Linguíʃtica: **And what are the points you consider most critical to traditional sociolinguistics?**

Tony McEnery: Well I think I wouldn't necessarily be critical of a traditional sociolinguistics. I enjoyed studying it as an undergraduate and I enjoy talking to sociolinguists, but I think that what the corpus potentially can offer to sociolinguists is the opportunity to move from small scale qualitative studies to large scale quantitative studies. This will allow them to see how general some of the findings that they have generated on the basis of small scale studies may be. That is not necessarily a criticism. It is more of an opportunity that I think exists for sociolinguists.

Revista Linguíʃtica: **Today, terms like cybercorpora have become more common. What are the advantages and disadvantages that you notice of compilation of corpora from the web?**

Tony McEnery: The advantages are very obvious. You can get data for a wide range of languages, lots of data from most of those languages and you can do so quite easily and at relatively low or little cost. So those are substantial advantages. The disadvantages I suppose depend upon your perspective. The web is a large collection of texts and a large collection of genres but the genres are not marked up explicitly on the web. So you are getting a large collection of material within which you know there is an important variable at work, genre, but you don't know what the genres are or even which documents are in which genre. It has been really interesting to spot those genres, the work by Douglas Biber, Mark Davies and Jessie Egbert, trying to actually organize what is on the web into genres, is fascinating. They did a fabulous project trying to do that but did not actually fully succeed in doing it. I think that project is fabulous because it shows that this is an incredibly

difficult thing to do even for humans, to actually work out what the genres on the web are and then organize texts into those genres. The web is an undifferentiated mass to some extent. Now, you can solve that problem by using humans to go to specific sites and say: 'well, this is news'. But if you are just doing it automatically, it is very difficult to organise these genres of the web for yourself. So that is an example of a real disadvantage I think, but you still need the curation of gathering and placing into categories of certain types of text when you are working with a webcorpora and doing so is principally a human activity at the moment. Actually the structure of a traditional corpus and its division into genres and relevant categories is an important distinction between a corpus like the BNC and something which is just ripped of the web. There is much more useful structure in the BNC than what you get when you just rip something off websites where you don't necessarily know what you are getting in terms of genre, for example.

Revista Linguíʃtica: **To what extent, in your opinion, the compilation and analysis of megacybercorpora (especially from virtual interactions) can contribute to studies aimed at speaker's linguistic competence? And for studies on language variation and change?**

Tony McEnery: Again you get the basic problem that you don't know what you're gathering from the web. Sometimes you get a lot of material and you don't know whether you are looking at expertly or professionally authored writing or something produced by someone who has a lower educational level and doesn't normally publish texts. So you get all sorts of things mixed together. In terms of looking at speaker linguistic competence, if you just look across the web, you get a wide range of speaking and linguistic competences. They will be present in your corpus but you don't necessarily know the origin of the variation. Is it education? Is it due to proficiency? Are they native speakers of English or are they actually second language speakers? Do they speak different varieties of English? You might think that one person isn't as competent as another but in fact what you see is that they speak different varieties of English and in their own country they are perfectly competent. If you judge Singaporian English as British English you might consider it deficient, but that›s actually a very unfair comparison. So again, in this undifferentiated mass of data that you might gather from the web sometimes, it might allow you to address these questions but it might also mean that you fundamentally misunderstand what you are doing or misrepresent the data. In terms of language variation and change, I actually think that the web materials are potentially most helpful, because, say, they may allow you to look at lexis changing in real time. It changes relatively rapidly. So you get new words developed and brought into the language, and, if you are using corpora while you are gathering them to look at language over 20 years time spans like with the Brown family corpora or the new BNC, you are going to miss a lot of words that come into existence and then disappear in that time frame. Something like the web does allow you to monitor and observe potential neologisms arising in a language, because that can happen at a terrifically rapid rate and similarly the death or obsolescence of words can happen at a terrifically rapid rate. So I think that is one example of an area where actually web based material can be very helpful indeed.

Revista Linguíʃtica: **Corpus Linguistics dialogues with various other branches of linguistic studies. It is observed, however, a certain 'noise' to the sub-areas that take into account the context and the social profile of the speakers. How has Corpus Linguistics been overcoming these challenges, considering areas such as pragmatics and discourse analysis?**

Tony McEnery: Ok. How relevant is Corpus Linguistics to pragmatics and discourse analysis? In terms of discourse analysis, my general view is that Corpus Linguistics has a great deal to offer to discourse analysis. Discourse analysts analyse texts in depth. So do corpus linguists - we are actually looking at the same type of material. I look at texts, they look at texts and here by texts I mean either orthographic transcriptions of speech or a written document. Actually all we are doing as best as we can is scaling up observations or taking large scales of observations and going down to the level of discourse analysts. So I would always say that we don't want to set aside the frameworks of discourse analysis. They are very helpful. What we need to do is actually mesh the quantitative and the qualitative together, so that we can bring the best of both worlds together. In terms of pragmatics, people talk about whether or not Corpus Linguistics can be helpful to pragmatics. My own belief is, yes it can and lots of people have demonstrated that it can be, such as Karin Aijmer for example. She has done a lot on discourse and pragmatics and clearly demonstrated again that you can take small scale qualitative observations and then use corpus data to take those observations to a broader scale. I think you always have to remember that what we are doing in Corpus Linguistics is nothing that is deeply unusual. We don't in many ways study language in very different ways than non corpus linguists would. For example, if someone says 'I've done a small scale qualitative study, here it is and you can see all my annotations on this text showing how I've analysed language and what roles I think people are playing in here'. Well that to me is just a form of corpus annotation, albeith they are saying that they've done their textual analysis on one text alone. Well, for me then the only challenge is replicating that analysis enough times so that you can have a large body of text and then you can use a computer to start to process those codes and make large scale observations. What we are often doing in Corpus Linguistics is taking what people do on a small scale with texts and then putting that to a larger scale. Now that larger scale can be difficult to achieve. Some of these small scale analyses are highly labour intensive. So, if you are doing an analysis which is very lengthy on a text and there is no way of alternating elements of it, then you need a lot of money to sit a lot of people down to do enough of those analyses to build a large corpus. But that doesn't say that in principle you could not do it. It is just slightly impractical because you don't have the cash or the time or the resources. So I would probably say that there is nothing I think in pragmatics that makes it impossible to pursue with Corpus Linguistics. But the types of analysis that they want at the moment haven't necessarily been automated or even semi-automated, so there would be a lot of manual labour to go into building corpora of sufficient size.

Revista Linguíʃtica: **Which variables, in your opinion, are more important for the construction of the identities of the subjects and the exploitation of contextual information of language usage?**

Tony McEnery: Well, in my own work, two very important techniques which I use are keywords and collocation. Keywords are when you contrast one set of texts, maybe one speaker, with another set of texts, or perhaps another speaker, and you look for words which are usually frequent or infrequent when compared together. So lexical profiling of that sort is very important as is collocation – how words associate or not with other words in their immediate region, maybe five words on either side of them. These are two very important techniques for me. And they, in terms of representation, typically reveal quite alot about how certain ideas or people or objects are constructed in language. In all of the work we have done in discourse analysis, keywords and collocation have been two important ways in which we have approached the question of construction.

Revista Linguíʃtica: **According to the methodology of Corpus Linguistics, how to establish a link between the structure of information, knowledge and technology?**

Tony McEnery: The interface with technology is important. Without computers this approach to the study of language is really impossible. We need the computer to sort, count and retrieve information. We need to be able to say how many times does this word occur and the computer rapidly goes through a billion of words and says 'this many times'. We then need the computer to manipulate data and perform various mathematical procedures. So, on that level, a very basic level, there is a very strong interaction with technology. We can,with a high degree of accuracy in an annotated corpus, say which part of speech certain words are, this word here is a noun, this word here is an adjective, etc, etc. You can do that automatically. So technology helps us. Similarly with other things like semantics, you can achieve automated analysis to some extent with some languages. But at the other end of the spectrum there are other things you can't do using computers, yet. That is when you get back to the problem we were talking about before where heavy manual analysis and annotation is required before you can start to get a computer to work those things out. So, say, for example, if I wanted to just find all metaphors in a text, somebody would have to go through and do the categorization of the metaphors in that text. Although I am sure some computer scientists would say that they could do it automatically, I have my doubts. So, there are still things that a computer can't do that linguists would like really. The influence of computing on this gets much weaker when you look at such questions, but nonetheless you can't do Corpus Linguistics really without a computer. So, even when we are manually encoding texts in order to help the computer reveal patterns, it is the computer eventually that will do the counting up and the processing of this data, even if the linguistic data has to be manually annotated.

Revista Linguíʃtica: **Corpus Linguistics broke out in the 80s and 90s with the popularity of individual computers and subsidized important research in several languages. How do you evaluate the current situation of Corpus Linguistics? What were the advances, which are the biggest challenges and what are the prospects for this branch of linguistics?**

Tony McEnery: I'd say the current situation of Corpus Linguistics is very strong. But it is also entering a slightly new phase. Approaches to analysing large text collections and coding big data are becoming quite popular. So, in linguistics, Corpus Linguistics is principally interested in exploring language, carrying forward linguistically informed analysis of some depth to a larger scale. We would not have been able to achieve this without the computer. So, Corpus Linguistics is popular and more and more people are using corpora. Some people are using corpora in ways the corpus linguists don't typically. But we should never forget what the true advantage of the Corpus Linguistics approach is it is linguistics. We are actually interested primarily in language, in the patterns of the language, in communication, in the description of language, we are looking for an explanation of why something is in the data. We are not principally interested in just using any mathematical trick to achieve a specific effect and that is what tends to happen over in big data - there is no deep understanding of how languages work necessarily. We are interested, or I

am interested, in psychological reality and the processing of language itself. I am not necessarily interested in building computer systems that can spot people's names or computer systems that can tell you if this brand of butter is liked on twitter or not liked on twitter. Though it is very interesting to the types of people that like to do such things, I am sure, I am not interested in that personally. I am interested in issues like saying how a group is represented, who has power or what are the consequences of people having power on language, how is that evidenced in the corpus, who uses bad language, why they are licensed to use it, what is the historical context in which this type of language use arises etc etc. These deep linguistic and broadly social questions around corpus data are what I am interested in. One of the biggest challenges actually is that Corpus Linguistics gets misrepresented as big data and we lose the reflection back on linguistic theory. Overall, However, the prospects for corpus linguisticsare very strong. 20 or 30 years ago not many linguists outside of the small circle of people who identified as corpus linguists used corpus data. Now it is quite common to find people in syntax or semantics, phonetics, discourse analysis, sociolinguistics, second language acquisition study using corpora - in the past that was unheard of. So, the prospects I think are very strong because more and more different types of linguists are using corpus data. Other academic subjects which study languages to some extent, sociologists, discourse analysts, have also started to use corpora: people in law use corpora now, recent court judgement in the States where they discuss the use of Corpus Linguistics as evidence in court, a whole range of people who have research questions which are principally focused on language who isn't just linguists, a whole range of research in the humanities and in social sciences who do that, are also starting to use corpora. So that is good too.