Open data in linguistics

Transcript of video

An interview with Professor Laura A. Janda, UiT The Arctic University of Norway

Date: December 2022

I am a linguist and a researcher. I do a lot of research on my own, and in the last 15 years or so, linguistics as a science has really changed a lot because we've gained access to huge quantities of data. And also we've gained access to very sophisticated software for analysing statistical tendencies. And we've really found it's been a theoretical sea change in our community, because a lot of things that we thought were very simple questions, where there were just yes and no answers and it was very clear line categorisation – we found that a lot of those really are statistical tendencies, and that there are many more factors at play and that they are much more complicated.

What experiences made you realise the importance of sharing data?

Around the year 2007 I went to a conference and I realised that I needed to learn statistics. And I went back to my University and took courses in statistics, and since then I've even written a textbook for linguists who want to use statistics, and I've developed a course here at UiT for linguists who want to use statistical methods in their research. And I realised that one of the hardest things about using or learning to use statistics is figuring out what model fits your data. It really helped me to see examples of what other people had done. Because, if I could see an example and I could say oh this is an example that's similar somehow to work that I've done, then I could say okay, I can relate to this. Because when I was starting, I had to take – I didn't take my courses in a linguistics department – I took my courses in a psychology department. And my teachers were psychology professors. Psychologists have been working with statistics much longer than we have, and so they're much further along on this learning curve.

But another experience that really pushed me in this direction is that I have for many, many years been the associate editor of our journal Cognitive Linguistics, and actually, recently I did a survey of all of the articles that have been published in that journal since it was founded in 1990 to the present. Our journal has always been data friendly, if you will. So there's never been an issue of our journal that's been published without any statistical analysis of data. However, around the year 2008, right around the time when I realised I needed to change, we also experienced the first time when we cross that 50% line – so 50% of the articles being published in our journal involve statistics – and we're probably never going back. I don't think we're going to go to 100%, but we're now very much

dominated by statistical analysis of our data. And I found also that it's a problem as an editor and as a reviewer – I also review for many other journals – if you don't have access to the data, if you can't see that data. So it's very, very important to provide access to that data so that others can see how it was done – learn from it. Others can try to replicate it. In that way we support the scientific method and really the integrity of our field overall. And it's also important just for transparency so that there will not be any fraud. We haven't had any big scandals in linguistics the way that we've seen perhaps in medicine and such, but you know it's always possible for people to fudge their data a little bit and it's harder to do if the data is all made open and public and it's all available. So those are some of the reasons why I got started, and then I felt like it would really help if we had one place, you know one-stop shopping, for linguists to find data and find the code and learn about it.

How are you involved with open data?

We got the idea of launching a website that would house those kinds of open data resources. And we went to our library and went to people in our library, and to our great delight they thought this was a wonderful project and were willing to spend months and even years in meetings with us, and they took care of so many of the sort of professional and technical sides of the question that would have been very difficult or really impossible for me to tackle on my own or even with my colleagues just here in linguistics. So this was really very much a partnership and we were very, very lucky that we had excellent colleagues in our library to help us out with this project.

I sort of – when you do a bunch of theoretical studies and statistical studies, after you've done a study and you've moved on to another study, and maybe – you know, a year or two later you want to go back and you want to reuse some of that data or you want to at least take some inspiration from it – sometimes it's hard to find your own data – even your own data! Or even understand how it was put together. Because, you know, all of those fields that you have in your in your files, if you haven't annotated them well enough. I mean, today of course I know exactly what all of those fields mean, but will I know in a month or in a year or in ten years? But the nice thing about having a resource like TROLLing is that it really forces me to upload all of my data in a place where I can find it again. And I can show other people where to find it. And also, if I've gone through the exercise of actually annotating it in a way that I hope makes it clear even for somebody who doesn't know me, and has no previous knowledge of my data, then hopefully it will be clear enough also for me when I go back to that data and look at again.

And it's become much better, and it's also way easier – again, for me it's easier to go back to TROLLing and find my own data, and find my own code, and I know it's always there and it's safe, than to have to dig around in my own files. And I use that also in teaching too, because I have a course – I have this book and I also have this – this is one that I worked on and this is also a text book that just came out like a year and a half ago, that I use in my course. I mean, she has some datasets

and analyses for people to go through on a website that goes with this book, but I have my own data and there's something different about your own data that you know it, right? And so I give my students my own dataset for each type of statistical analysis they're supposed to learn. I give them my own dataset and my own code, and then we work through it. I can answer, you know, all of their questions and really give them a full experience of what it's like to work with your data and your code.

So it's another thing that – it's kind of like a myth, I guess, that I had to break free of in order to move into this new way of doing linguistics. Because it's not like you can just collect data and then shovel it over to some statistician. Because the statistician once again – you know you say the word 'verb' and the shutter is going down. And they don't understand anymore. I mean you really have to just analyse the data yourself because the statistician will never understand it the way that you do. And also, you have to have some idea of what the models are that you're going to use at the end, in order to collect the data that will be amenable to that kind of modelling and that kind of analysis.

One of my colleagues said, when we were making the instructional videos, he said: "Laura you have to make these instructional videos such that even your grandmother could upload data in TROLLing". And I think we came pretty close to that. I think it's pretty self-explanatory with the instructional videos. And I've always felt that research and teaching go hand-in-hand. I mean, I've never been involved in a research project that didn't have some sort of a teaching angle to it. And conversely whenever I'm teaching I am always trying to think about, you know, what do we still need to learn. And that's one of the great things about teaching, that you then see the students – you can see those gears turning in those heads. And you can see that they see it from a different perspective. They come up against a wall, they come up against a problem. They don't understand why something works this way. And then you say, oh we need a better explanation or we need to learn more about this phenomenon. So I learn constantly from the students and that feeds back into the teaching and research. And so I think that they just go on – it's a continuous cycle.

They're sort of like getting a simulated experience of hands-on working with the data. They get the data, they get the code, we go through it, and we all will sit there together – they all have their computers open. But yeah, it's like a hands-on experience of working directly with the data.

What do you consider to be Open Data concerns?

I have another thing actually I could say also about dangers, if that's interesting. Because there's one thing that has concerned me quite a bit recently. I mean, we have a challenge sometimes finding academic research positions for many of our graduates in linguistics. However, there are some corporations that are very interested in hiring statistically capable linguistics graduates. And these are mostly big corporations like Google and Amazon and Apple and Facebook and such. And these are the public ones, right, in the sense that everybody knows that they exist. But they are doing a lot

of clandestine research on you and me using linguistics and using big data. And everything that they do is kept under cover. That is all company secrets. And some of the type of … I mean, it's spyware. Let us put it that way. I mean, they are spying on us, they are using linguistics and data techniques in order to spy on us. And they are not alone. I mean, there are also various governmental organisations that are doing similar things. And spyware operations – and this is something that's pretty much unstoppable. I mean, it is just going to happen. We can't prevent it. But the more that we put things out there ourselves and make things as public as possible, I think that's our only defence. That we have all of these things in plain sight, and don't let it all be shut behind the doors of spying operations and major corporations.

What inspires you and makes you optimistic about the future of Open Science?

So when you ask about the future. My hope is, well I think that the statistical studies and data studies in linguistics are here to stay. I mean, I think that's definitely a part of our future. I think that in the future probably all linguistics programs will have courses in statistics for students. And that will be part of the expectations of submitting articles to journals. So, my hope for the future is that TROLLing, our website, will continue to be a clearinghouse for those materials. A place where people can upload the materials, and also share with each other and learn from each other. One never knows when one collects data what sort of structure in that data might have been overlooked that somebody else could find. And that's one of the really exciting things about this time that we're living in is that suddenly we have access to so much data and a way to look for the structure, thanks to the sophisticated statistical software. So absolutely, I think we're living in very exciting times in that sense.

I can maybe name something really recent. I wonder if I have it here – yeah, here it is. So this is a dissertation that was defended in Leiden. I actually had met the author at a couple of conferences before, so I knew approximately what he was working on. And he knew something about what I was working on. But then I was asked to be an opponent, to be an examiner at his dissertation defence. And so I got a copy and I was reading through it and then I realised that he had taken the method that we had used. And he had gotten it from TROLLing, from our open data site. He had taken that method and used it on a different dataset and in a different way, and it was just so exciting. I practically cried. I mean, it was just a really, really exciting moment. So, come to think of it, this is the stuff from … these are actually pictures from an article that I wrote together with a colleague of mine about modern Russian.

And this is the kind of thing that can happen. And this wouldn't have happened if it wasn't for TROLLing, because he might have read my article, but then he probably would have said 'Well I don't know how to do this and how am I going to figure it out and everything'. But he just did this without even having to call me or ask me or anything, you know. He just went to TROLLing and downloaded it

and saw how it was done, and said ‘Oh yeah, I can do the same’. And did the same and wrote his dissertation.

What still needs to be done to get more people to share and open up their research data?

I think what a big challenge it is to educate people so that they understand that everybody gains but nobody really loses anything. And that's one of the things that we have also safeguarded in TROLLing. Because we have instructions for how to cite the data, and once you put up your data in TROLLing, then everybody will only know that that was your data because your name was on it first, and we have the posting dates and all of that information. And you can't lose anything. All you can gain really is more perspectives from more researchers and maybe more interest in your research.

Finally, could you mention one important positive consequence of data sharing?

One thing about linguistics: As I mentioned, in psychology they've been doing statistical analysis for a long time, and we've come to this rather late. But that means that we're in this conformity period where we're really discovering what are the methods that are going to work best for us. And by sharing our data and doing this in a very sort of open public community type fashion, we can really decide our best practices in our field. And really help our whole field move forward by setting standards. And I think that's also really important.