FROM LANGUAGE RESOURCES TO LANGUAGE COMMENTARY: THE INTERVIEW WITH BEN ZIMMER

Iztok KOSEM

University of Ljubljana, Faculty of Arts

Ben Zimmer is a linguist, lexicographer, and all-around word nut. He is the language columnist for The Wall Street Journal and former columnist for The Boston Globe and The New York Times Magazine. He is the recipient of the first ever Linguistics Journalism Award.

Ben Zimmer has worked as the executive editor of Vocabulary.com and the Visual Thesaurus. He was also editor for American dictionaries at Oxford University Press and as a consultant to the Oxford English Dictionary.

This interview took place at eLex 2017, a biannual conference on electronic lexicography, where Ben Zimmer gave a keynote talk titled »Defining the Digital Dictionary: How to Build More Useful Online Lexical References«. The interview was conducted by Dr. Iztok Kosem from the University of Ljubljana (Faculty of Arts and Centre for Language Resources and Technologies).

You call yourself as "an all-around word nut". Could you elaborate on that?

I can call myself a linguist and a lexicographer, but it's a way of describing the fact that I approach the matters involving language and words with a great deal of passion. From an early age I was paging through unabridged dictionaries and was fascinated by the history of the English language and the journeys that words take. That love of language has stuck with me.

Nowadays, I get to do what I love, which is to think about words, think about language and I try to communicate that to a broad audience, which I get to do through the language commentary, like for example I'm currently writing a weekly column about language for the Wall Street Journal.

Which linguist or linguistic theory has influenced you and your work the most?

I went through different phases. I started off studying formal linguistics as a linguistics major at Yale University, where there's a strong tradition of structural linguistics, which could very much be seen as going back to Saussure and Bloomfield, who are the progenitors of structural linguistics. But I soon found myself very interested in language in its cultural context, so I moved into more of a sociolinguistic line. I pursued that at the University of Chicago for graduate school. I was exposed to all sorts of different thinkers about language, someone like Roman Jakobson for instance, and the thinking of functional approaches to language, which was helpful in terms of thinking about what language can accomplish beyond just denotation.

Another scholar who I was drawn to in graduate school was Bakhtin, who was a theorist that I found very intriguing in terms of his approach to language and his ideas about the concept that gets translated into English as heteroglossia. He's talking about the centripetal forces that try to keep language together, which could be institutional forces that treat language as a unitary whole. And then you have these other forces, the centrifugal forces, that lead to not just one

unitary voice but all these different voices. Even though Bakhtin was doing it through literature, you could apply that to language in general and think about the many voices that language is shot through. To think about language in this new way was just exciting to me, because it went far beyond what seemed to me a drier picture that you would get from a more formal linguistic approach.

When did you start working for OUP (Oxford University Press)?

OUP was where I ended up after deciding that I was done with academia and that I wanted to again return to that passion for lexicography from when I was younger. At that point I was already very interested in what technology can do for us in terms of painting a broader picture of language and specifically lexical items and their history. Even before I started working at OUP, I was in touch with the editors of the Oxford English Dictionary (OED). I started working as a consultant on various aspects of the history of American English terms, American slang and some other areas that seemed underrepresented in the OED and I felt that I could help with. The newly digitized databases that were coming online allowed us to do targeted searches for particular words and phrases for the purposes of antedating, i.e. finding earlier examples in print than what the OED currently had, but also to paint a more accurate picture of the life of a word.

Then there were the efforts that OUP was making at the time to expand online. Oxford dictionaries online was bringing together the US side with the New Oxford American Dictionary and the UK side with the Oxford Dictionary of English. That was my first real exposure to corpus-based lexicography. The Oxford English corpus was being developed at the time and even though I was not on the technical side of that project, it was a key part of our work. It was really eye-opening for me to see what corpus-based analysis could do. We were also using the corpus for other projects, for instance the second edition of the Oxford American writers' thesaurus, which has various features that are intended to help writers think about word choice, beyond just simply giving a

list of synonyms. One thing that we developed for that was a feature which allowed you to see collocational patterns extracted with the Sketch Engine. In a thesaurus, you might see a list of synonyms for *weird* or *strange* that might include *eccentric*, but rather than simply thinking of these as interchangeable words, you can use the corpus to look at what types of nouns are typically modified by these adjectives. For example the adjective *eccentric* very often may modify someone rich, a millionaire, a billionaire, or some relative, like an uncle. We built that into the Oxford American writers' thesaurus around 2007-2008.

It was exciting to be part of that, thinking how these tools can go far beyond what our native speaker intuition may reveal about language, and understanding usage on a broad level too. No matter how well-read you are, you can never have a perfect understanding, a perfect *sprachgefühl*. Corpus techniques help to provide insights that then allow you to understand issues like the nuances of meaning of words that are ostensibly synonyms, or the diachronic changes that would happen that would require you to revise an entry, for instance a word like *browse*. Which may originally started with animals just nibbling on leaves, then got extended to passing over books and of course nowadays applies to searching on the internet and the world wide web. Corpus techniques let you understand how that word works in the world and what other words it likes to hang out with, for example, do you *browse over,* is there some preposition that's needed? Or is it just used transitively? I was intrigued by all of these things and I also got a feeling at the time that I wanted to be able to communicate these ideas broadly, not simply use it to make better dictionaries. The latter is good but I wanted to talk to as general an audience as I could about these things and explain what we can learn about the history of words and the way they work, how they developed diachronically and also synchronically, at any given time.

I had to kind of unlearn some of my habits of academic writing in order to write for a more general audience. While I was working at OUP, I started contributing

to Language Log, which is a group blog that was set up by Mark Lieberman, a linguist at the University of Pennsylvania, and Geoffrey Pullum, who's at the University of Edinburgh. That was some good practice for me in terms of learning to write for a more general audience. At the same time you had the freedom to get into a deep dive into linguistic issues as much as you wanted. So it gave me the sort of the freedom to think about these terms beyond just what I might write in an academic paper, and also to think about language issues that come up in the news in politics and popular culture. How a linguistically informed perspective can enrich people's understanding of some aspect of language that is worth commenting on, whether that came from politicians, music or wherever else.

Then there was an opening with a company called Thinkmap, which at the time was known for the Visual Thesaurus they had developed. They were looking for someone who would oversee both lexicographical projects and content on the site, since they had a kind of a magazine for people who subscribed to the Visual Thesaurus. It was something for word lovers, the sort of content that you would want for people to feel like they were part of this community. They could be writers or other people who work with language, or just word lovers in general.

A perfect fit for you.

Right, it felt like it hit all of the things that I wanted to be in at that point. It involved some very interesting cutting edge technology in terms of the visualization of linguistic data, while it also required building a community through interesting writing about language and about words. They gave me the freedom to shape that how I wanted to. I was doing a lot of my own writing as well, so I started a regular column for the Visual Thesaurus called Word Routes and got various people to write for me. It was a very different type of culture from OUP, where like with any big publisher, decisions may happen very slowly over time and things have to be run through committees and so on. When you get into the tech culture and startups, you meet small groups of people who are

committed to making something happen. If you have a good idea and it is embraced by the people you're working with, you can actually make things happen very quickly.

One of the fun projects that we had at the Visual Thesaurus was to make a spelling bee, which was basically an online spelling quiz. We had high quality audio pronunciations for all the words in the Visual Thesaurus and we created this adaptive system which involved you hearing the word and seeing its definition, and if you typed in the word correctly, then it gave you a harder word and you had to progress through it. So it would adapt to whatever your level was. We started thinking about what else we can do that takes this adaptive approach. That led to the creation of Vocabulary.com. We were talking to teachers and tried to understand what their needs were, because there really didn't seem to be a good website for this sort of thing. Vocabulary instruction in the US at least was dominated by a kind of flashcard approach involving rote memorization.

Is vocabulary instruction a big thing in the US?

Yes, a lot of it is. But there's been a move towards not just simply vocabulary that you have to memorize, but understanding vocabulary in context as a part of reading comprehension. More and more, the tests in the US like the SAT are taking that approach. So that was also something we were keeping in mind with Vocabulary.com. You could build an adaptive vocabulary quiz which was basically just the flashcard approach, but what else could you do with it?

That's where some of the corpus-based approaches that I was interested in could be applied effectively. We ended up building up a kind of a hybrid corpus that involved everything from news feeds to literary sources. Originally, just to get started, the literary sources was public domain literature, like Project Gutenberg, just because that's an easy data set to start with. But since then it's grown quite a bit and we were able to incorporate books that were often taught in school curriculum, as well as recent fiction and nonfiction that we could scan

and digitize and then extract sentences from. Those sentences could then be used as example sentences in the Vocabulary.com dictionary, as well as, with some human intervention, become the basis for some vocabulary questions where you have a decent example sentence for a particular word sense; we were always cognizant of the fact that words can be polysemous and that learning how a word works involves mastery of all of its different meanings. Of course that involved projects with word sense disambiguation, e.g. taking a sentence from the corpus, matching it with a sense of the word and then using that as the basis for a question. You could do it by requiring to fill in the blanks, i.e. asking which word fits the slot, or you could ask what a certain word means in a certain sentence.

So you had to not only build a corpus but compile a dictionary as well.

Both Visual Thesaurus and Vocabulary.com started with WordNet as the basis. At its heart, Visual Thesaurus was basically a visualization of the semantic relationships in WordNet. When we started Vocabulary.com, we started from there but realized that WordNet was never designed to be used as a dictionary. Whatever we did with WordNet would require a lot of extra work in terms of making it useful to people. It was very clear that for this to have value, we needed to create our own original work and not just be piggybacking on WordNet.

Another thing that became very clear from talking to teachers and other people in the field was that dictionary definitions very often do very little for students' understanding of words. Especially when they're written in that sort of dictionary style. You can have an adjective that starts off with something like: »of or pertaining to…«

A lot of us remember an English teacher that really made an impression on us and got us excited about literature and that sort of thing. We thought about how would that really good teacher explain the meaning of a word to you. That's

obviously very different from just looking up the definition of a word in a dictionary. That led us to write more conversational or discursive explanations of words and their meanings. Through various techniques we identified what the top vocabulary words were and did a massive project just to cover these words first. At first it was five thousand words and then ten thousand and fifteen thousand... This resulted in Vocabulary.com having a kind of a casual voice talking to you about words and its meanings. We felt that was something that would connect with students much more than just a traditional dictionary approach.

Could you give our readers an example of this?

One example that I like is the word *hirsute*. A dictionary definition would just be »having or covered with hair«. But then if you get this more playful approach – "What do Santa Claus, Bigfoot and unicorns have in common, aside from the fact that they're completely real? They're also hirsute, very very hairy creatures." When you read something like that, it's playful and engaging, it has a little edge to it that you don't necessarily expect from a reference work. It's having fun with the language and injecting humor as well. That approach really seems to help people who are learning vocabulary, whether it's students or older people. It cements the meaning of the word better than it would if you were just looking at a flashcard.

Did you get any feedback on this?

There was plenty of feedback because this is a case where we wanted as much feedback as possible, especially from teachers in schools. The way that the business model worked, we were able to keep the site relatively ad-free, not by having a subscription model which we use for the Visual Thesaurus, but by offering a premium edition to schools and school districts. It would allow teachers to track student progress, and create vocabulary lists for their students that might be based on a particular work of literature, for instance.

One thing that we came up with for Visual Thesaurus and ended up using in

Vocabulary.com was called VocabGrabber, which allows you to input any amount of text, identifies what the vocabulary words are and then creates learnerable lists out of that. All of those things make teachers' jobs easier. They don't have to spend their time trying to come up with vocabulary lists based on chapter one of their book, they can let Vocabulary.com do that for them. This frees up teachers to do what they need to do.

Every interaction that the user had helped us understand more about a word and its meanings, in the sense of which ones are most difficult. Sometimes there can be a word that has a very straightforward meaning and one that's more difficult, so you have to always be thinking of the full gamut of meanings of a word. And by making Vocabulary.com adaptive so that the user is actually being fit onto a curve, basically, then that allows the vocabulary questions to be given that will be appropriate.

We were also able to calibrate what difficulty do we want a certain person, based on their level, to be getting. For example, fifty percent right and fifty percent wrong turned out to be a little too discouraging. You want people to get more things right than wrong, so you turn it up to maybe 60 or 65 percent right. That way they're still feeling challenged because they're still being presented with vocabulary that is challenging to them and that they feel they need to learn and master. With this adaptive approach, the more data you have the better, so this is very much a big data approach.

One other thing about Vocabulary.com is the fact that it's topical. In education I think that's quite important. When a teacher gets into a classroom, sometimes the best motivation is to talk about something that's currently happening or has happened recently.

Absolutely. Beyond just its use in English language arts, we got immediate pickup from social studies teachers, who might be talking about current events and relating that to history and that sort of thing. Vocabulary.com can take a news article and create a learnable vocabulary list, so that students know what

to expect when they have to read an article in for example the New York Times, which does have vocabulary that might be more advanced. But that's the kind of language that they're going to need in order to really make sense of world events.

How do you deal with new meanings of words?

That was something which during the time I was there we weren't able to address yet. It was sort of left for the future. Because of the approach that we took and because of the fact that I was the only lexicographer on staff, there was really no way to make updates the way that major dictionary sites do.

With Vocabulary.com, it was never a goal to try to keep up with the latest changes in the language. I was doing that myself through my other activities, for example as a member of the American Dialect Society and the chair of their New Words Committee. A lot goes into the selection of the word of the year, which involves keeping tabs on neologisms. Then I took over the feature in the journal American Speech called Among the New Words, which has been going on since 1941 and all of this time has had the mission of focusing on lexical items that are coming up before the major dictionaries can get to them. Nowadays though, whether it's Oxford, Merriam-Webster, Cambridge or American Heritage, they're much more agile and respond more quickly, because they do updates every six months or on whatever timeframe they want.

But through my work with American Dialect Society and Among the New Words, as well as through my writing for various language columns, like the ones in the New York Times, the Boston Globe and now The Wall Street Journal, I was interested in talking about neologisms and also old words that become prominent in a new way. Through all of that I was constantly needing to keep tabs on the latest things that were happening in terms of the English lexicon. So I kept myself quite busy with various other things concerning the latest developments in language.

One of the things that people have probably started to expect since this shift from print to electronic medium has happened, especially online, is an immediate response. For example, when something happens, you can see the news everywhere. Everything is discussed and you get information quickly. If the dictionaries are not reacting that fast, that's where the users end up being disappointed. How do you see lexicography reacting to something like this? For example, European survey on dictionary use has shown that while reliability is the most valued dictionary characteristic, ease of access, free access, and up-to-date content are also high on the list.

In practice this can be difficult for online dictionaries to achieve, as I've discussed in various places. Dictionaries traditionally have not been agile enough to make these changes and there are institutional forces that keep things going at a more conservative pace. The world that we're living in now and the way that people expect to get information about everything, including language, requires a much nimbler approach and the ability to react quickly to things that are happening in current events, politics, popular culture, and so forth. In the US, we see dictionaries approaching that, Merriam-Webster for instance has gotten attention for its use of social media, so that when Donald Trump has some use of language that people are talking about, Merriam-Webster, which by many Americans is sort of seen as *the* dictionary, will then come up with a response. But really all they are doing is something they've done for a while. They are keeping track of what people are looking up in the dictionary, noticing what's spiking and then explaining what the interest is and where it's coming from. There might be something in the news that is driving people to look up a particular word, even if it's not a difficult one, not a vocabulary word.

As an example, at the beginning of 2017, after Donald Trump's inauguration when people were fighting over the attending crowd size, his spokesperson Kellyanne Conway gave an interview where she used the phrase *alternative*

facts and people were up in arms about that. Merriam-Webster noticed that people were looking up the word *fact* in order to try to figure out whether there really could be such a thing as an alternative fact. So they wrote a Trend Watch piece about that in which they acknowledged this development as based on what Kellyan Conway had said. When they put that out on social media, it became this thing where *the* dictionary was somehow in conflict with Kellyan Conway. People made a big deal out of it simply because it sounded like it was sort of this voice of authority – the dictionary, combating what she was saying about alternative facts. It's been fascinating to watch how people have been engaging with dictionaries in a new way. So that shows that beyond simply just providing definitions for words, there is a role for a kind of knowledgeable voice about words and how they work in the world that people look to when they're uncertain about things. Very often their uncertainty congeals about a real meaning of a word. If there are debates about marriage because of same-sex marriage, it's actually a question of the *definition* of marriage.

Your career has now moved on to actually being on the front line, so to speak, of communicating directly with dictionary users, as well as with a wider audience.

I think it's an exciting time for language commentary because it used to be, at least in the US, that the people who wrote about language didn't necessarily have any background in linguistics or lexicography. They were very often just journalists who decided to write about language. Now there are all sorts of great new linguistics podcasts. There's one that I've been involved with called Lexicon Valley that's done by Slate. For a while they had me on to present a mystery word and then sort of tell the story of its history. So there are all sorts of new possibilities to get a more linguistically informed sense of language out there and into the public. And there's a an appetite for it. People are looking for that kind of reliable knowledge. When I'm writing a column about a word in the news, I feel like I am engaging in a kind of lexicography, even though it's not writing a dictionary entry but it's more of a narrative lexicography. You use a

column of 500 or 800 words to bring the reader through from when a word might have entered the language and how it shifted over time. Doing that I actually use the same tools that I would use for some research for the OED. Also sometimes I'm able to talk about some very interesting discoveries that I've made. An example of one that I'm proud of would be the usage of the title for a woman, Ms., regardless of her marital status. This is something that took off in the 1970s because of the feminist movement, but it had been proposed at various times before that. I was able to find that the original proposal appeared in an American newspaper from Springfield, Massachusetts, the Springfield Republican, in 1901.

There was another thing I wrote about for the Wall Street Journal, which wasn't my own discovery, but was found out by an OED researcher – the origin of OMG as an abbreviation for "Oh my god". It appeared 100 years ago in a letter written by a retired British admiral to Winston Churchill, where he was poking fun at the various orders of knighthood and suggested a new one, OMG, standing for Oh my god. Winston Churchill didn't pick up on that and nobody started really using it in a serious way until the mid 1990s when this sort of initialisms like OMG, LOL and all the rest started catching on. This particular example from 1917 was found thanks to Google Books. It's things like Google Books and these other digitized databases that allow us to tell these stories in a way that was impossible before.

In your biography I read that you received the first-ever linguistics journalism award. How do you see this award?

It was a very nice surprise. The Linguistic Society of America created this award and it fit very well with what I was doing, so I got the first one. There have been a few winners since then who are doing it in their own way, there are indeed many different ways to do it, from long form journalism to more sort of in-depth explorations of things. My columns for the Wall Street Journal tend to be on the short side, but then I also write longer pieces in various outlets like Slate.

My work combines the research practices that I've learned through lexicography with journalistic practice. Sometimes to tell the story in the right way, you have to get on the phone with the person who popularized the word, that is if you can actually identify that person. Usually you can't, but sometimes you can find out that there was a key person who is alive and you can talk to. This way you get actual first-person account of how a word came into the mainstream. Language commentary has been traditionally focused on particular usage peeves, things like how people use the word *literally* or something like that. It's such a tiny portion of the language. People get irritated by the way that others use a particular word or phrase, but there's so much more out there. I've been doing this for years now, I'm writing columns every week and I'm still in no danger of running out. That's what's exciting. It feels like every word has its own story. I've been able to tell a lot of them, but there's still no end to it.

In one of the past eLex conferences, there was a roundtable with a provocative title "Will there still be dictionaries in 2030?" The conclusion then was that there will be, but they won't be called dictionaries anymore. What do you think of that statement today?

When it comes to language, it always seems impossible to predict anything. In this case it's a combination of language and technology. I think that despite the fact that dictionary publishers are hurting financially, it's very difficult to keep staff together these days to publish dictionaries in print and that the online business models haven't been worked out, it feels like there's more interest than ever in the things that a dictionary can provide.

We used to do print dictionaries and now we translated that into the online space. Even just in the past five years, we've had all these different shifts and people are just trying to keep up as best as they can to make things work. There are imperatives now that lexicographers never had to worry about in the past, like making Google happy and making sure that you're high up enough in the

search results. Because of that uncertainty, it does feel like dictionaries as we know them, even the online dictionaries, which are still representations of the old print model, may not actually be what we call a dictionary in the future. I think there will be something that we call a dictionary but it could take many different forms.

In many European countries, mainly the ones with a small number of native speakers but also a few larger ones, dictionaries are publicly funded. There is also a big push for open access. What is your view on this?

In a small European country that has a stake in making sure its national language thrives and has those kinds of resources and makes those funds available so that these things can be shared in an open way, it's a very different model from what we have in the US. I don't foresee a time when the US would be able to follow that model, just because sadly that's not the way publicly funded scholarship works in the US. If anything, it seems like the trend is going in the opposite direction and we won't be able to rely on state funding even for such great projects like The Dictionary of American Regional English, which has been going for more than five decades but is shutting down this year. It was actually getting funding from the National Endowment for the Humanities, so the type of funding that used to be available for grand projects like that seems to be going away as well. It's difficult because people now expect to have open access. Dictionary publishers recognize that there's no sense in putting up a paywall for your dictionary content, but then if you're just giving it away for free, you're stuck with the problem of generating revenue, whether it's from ads or other sources.

Of course I would love to see more collaborative approaches. Back when I started doing lexicography and was consulting for the OED, I was always fascinated because they were doing crowdsourcing long before it was called crowdsourcing. Volunteer readers have been pitching in for the OED from the

19th century onwards, just simply because this was a great project that you wanted to contribute to. I was interested in how we can tap into that in the era when you can find online communities of people who are interested in particular things. One thing that the OED did for instance was a science fiction project, where they focused on words that have come out of science fiction. There's sort of a ready-made audience right there of people who have all the old science fiction, books and magazines and can help us hunt down where *cyborg* first appeared, for example. I was interested in extending that model to other areas of interests, like getting the baseball people looking into baseball and getting the jazz people looking into jazz and so on. Some are are interested in movies, others are interested in old cars. Whatever their interests might be, you can have those people contribute in a way that you couldn't do before. I'm still interested in what can be done in that kind of collaborative fashion, where it is as open as possible and you build towards the common goal. Sadly, the problem of funding keeps coming up.

How would you approach making a completely new dictionary if you didn't have any limitations imposed by a business model, in terms of considering the users, their needs and their appetites, as you put it?

I like a very open-ended approach because as Erin McKean has been saying with Wordnik, we have these sort of artificial boundaries or barriers to wordhood which may not really make any more sense. We don't necessarily have to worry about whether a word is still going to be in use ten years down the road. It's something that matters now and if it doesn't matter next year, then that's fine. It was something that was important and relevant to people at a particular time. I'd be in favor of giving each new word the kind of resources that are necessary to understand where it came from, how it's getting used, how it might be shifting in terms of its usage and that sort of thing. I think that would also allow people to think about language in a new way, not as a static thing but something that's constantly evolving and is a democratic process. It would be a kind of

grassroots effort from the bottom up, not some authority figure telling you what is a word and what isn't.

That does present unique challenges because if you recognize language as a constantly changing phenomenon, how do you ever really get a hold on it? You can't, but you can track the currents that are going on, and even if those currents may peter out or develop into something brand new, at least you can appreciate that people themselves through their interactions are the ones who are creating language together. That has been my ongoing interest in everything that I've done.

I think that's a perfect way to end it. Thank you very much for this interesting interview.

OD JEZIKOVNIH VIROV DO KOLUMN O JEZIKU: INTERVJU Z BENOM ZIMMERJEM

Ben Zimmer je jezikoslovec, leksikograf in velik ljubitelj besed in jezika. Piše kolumno o jeziku za The Wall Street Journal, v preteklosti je podobno kolumno pisal za časopisa The Boston Globe and The New York Times Magazine. Je tudi prvi prejemnik nagrade, namenjene izjemnim novinarjem, ki pišejo o jeziku (angl. Linguistics Journalism Award).

Ben Zimmer je bil glavni urednik portalov Vocabulary.com in Visual Thesaurus. Bil je tudi urednik ameriških slovarjev založbe Oxford University Press in svetovalec pri Oxford English Dictionary.

Intervju je bil opravljen med konkferenco o e-leksikografiji eLex 2017, na kateri je bil Ben Zimmer vabljeni govorec in je imel predavanje na temo »Defining the Digital Dictionary: How to Build More Useful Online Lexical References« (Opredelitev digitalnega slovarja: kako izdelati bolj uporabne spletne leksikalne referenčne vire). Intervju je izvedel dr. Iztok Kosem z Univerze v Ljubljani (Filozofska fakulteta in Center za jezikovne vire in tehnologije). Intervju se dotika različnih obdobij leksikografskega oz. jezikoslovnega delovanja Bena Zimmerja, od začetkov pri založbi Oxford University Press in ključne vloge pri ustvarjanju portala Vocabulary.com, do pisanja kolumn o jeziku za različne znane ameriške časopise.