Expert interview

Case: Wikipedia
Interviewee (IE): Amir Aharoni
Interviewer (IR): Anja Ebersbach
Date: 2020-03-20, 11:00am
Location: Skype

[Welcoming]

1 IR: So let's start, how long have you been working with Wikipedia? #00:02:32#

2 IE: Ok, so I started as a volunteer in late 2004 and it was very easy for me to start. My impression is that a lot of experienced wikipedians who have an experience of several years, they don't remember very well how they did start and what difficulties they had. I do remember and remember that it was mostly easy for me and I know why was it easy for me. That's because I was really familiar with the basic concepts behind Wikipedia, I didn't need an introduction for that. I knew what a wiki is. Because, for the most of people Wikipedia is the first wiki, and the only wiki they know, but I was familiar with the concept of the wiki even before I found Wikipedia. I stumbled upon some small wikis sometime in 1997 or 98 so I knew. #00:02:41#

3 IR: Really? #00:03:55#

4 IE: Yeah, I was learning software development and I was searching for some websites about Java programming or something like that, and I found one website that looked useful, and it was a wiki. So, it was weird at first that I could change pages on the website and I didn't do it much. I did not contribute very much to that website but thanks to that website I became familiar with the concept of the wiki. So, I was familiar with the wiki, and I was familiar with the concept of free software and free licenses. So, when I saw Wikipedia, the free encyclopedia, it immediately made sense to me. I knew about encyclopedias, I have loved encyclopedias since I was maybe six years old, and I knew what a wiki is, and I understood what a free license is, so it immediately made sense to me, that is why it was so easy. And I started in English and took me a few… I don't know, I saw quite quickly there are other languages, but it took me actually a few months to start contributing to Hebrew, and I also started contributing to Russian occasionally - Russian is my first, my native language - but I contribute much more to Hebrew, simply because… I know Hebrew perfectly and fluently. I have been speaking it for 30 years, and… it's just that the Hebrew Wikipedia community is very active, but it's much smaller than Russian, so it needs my help more, I can make a bigger impact in a smaller community, I think. So that's why I contribute mostly to Hebrew, occasionally to English, occasionally to Russian. #00:03:56#

5 IR: What is your favorite one? #00:06:01#

6 IE: Probably Hebrew, that's what I contribute the most. #00:06:44#

7 IR: Good. And the Russian was much bigger, when you started in 2004, than the Hebrew one? It still probably is…. #00:06:10#

8 IE: Of course, of course. Russian is like 200 million people who speak Russian and something like 8 million or 9 million people who speak Hebrew. For a relatively small language, Hebrew is very active. For a language with this number of speakers. But it's much smaller than Russian. #00:06:20#

9 IR: And which roads did you take on? You wrote, of course, as an author. #00:06:40#

10 IE: Yeah, I created some articles and topics that were missing. Like originally, at first, I mostly wrote articles about music that I like, about languages, because I loved linguistics for many years. I studied linguistics in university, so I wrote a few articles about languages that were missing, like I remember that I wrote about Latvian, about Gaelic, and about few others the articles on Hebrew Wikipedia. I corrected some articles and I also, like, I was in university when I started and sometimes later, I completed my BA in linguistics and Hebrew language. So, I can now say that I know Hebrew professionally. I have a degree, so I know the spelling rules very well, like much better than average Israeli. I can correct spelling quite well, so very often I just search for common spelling mistakes and I correct them. #00:06:47#

11 IR: Is it an issue in Hebrew? Is it very hard to write? #00:07:55#

12 IE: Yes, for many reasons. Because here, spelling is very strange, it's a very ancient language and spelling is very traditional. There are lots of weird rules that a lot of people don't really know. People usually write according to some intuition, which works most of the time, most people do not care very much but it is an encyclopedia so it's supposed to be written in very precise, academic, standard language. There are few people, including myself, who can take upon themselves to improve the spelling stuff. #00:08:01#

13 IR: Ok. But you have never been administrator or something like that? #00:08:47#

14 IE: Oh, yes. I was an admin. I have administrator rights in English and Hebrew. I do not use it very often. A lot of administrators spend a lot of their time searching for vandalism and blocking people and reverting and stuff like that. I do not do it very often. I can try for maybe once a day, for a few minutes a day, I just try to take a look if there is any recent vandalism, any request to block somebody and something like that, but I don't do that very often. Occasionally I like to help delete pages, restore pages, protect, but not very much. Most administrators in the Hebrew Wikipedia do it more than I do it. #00:08:52#

15 IR: As I remember you are part of the language team? #00:09:44#

16 IE: Yes, I was just a usual contributor to the Hebrew Wikipedia for several years. And after some time, I started going to meetups, like at least two meetups a year, in Israel. It's not a large country, but we have two meetups, sometimes more than two. I started going to these meeting people and after some time I joined the chapter as a member. The chapter in Israel, Wikimedia Israel. And then I was surprised, it took everybody by surprise that Wikipedia Israel was granted the

right to host Wikimania 2011. In Haifa. So, I was volunteering there, helping out with the organization, the conference. And some people from the Foundation approached me and just offered me a job, which was perfect because I was looking for a job. Since then, I am on the staff of the Foundation. #00:09:55#

17 IR: So, you are a professional there. I didn't know that. I thought you were a volunteer. #00:11:09#

18 IE: I do a lot of things as a volunteer which are things that are not part of my Wikimedia Foundation job. But I also have this as my day job, full time. There is the language team, also known as the language engineering team in the Wikimedia Foundation. We develop software. I can speak about this a little bit more. And there is also the language committee which is a separate thing as most members of the language committee are volunteers. The language committee does pretty different things, they mostly approve the creation of new domains, new domains for new languages. So, this is a volunteer committee. #00:11:17#

19 IR: Yes. We will come to this later, I think. As you have seen I have tried to cluster the questions. My first cluster is about content. From your perspective what do you think, how do the Wikipedias differ – apart from the numbers? #00:12:12#

20 IE: Not very much, I have to say. In the terms why the people write, what are the people's motivations… they are pretty much the same everywhere. And I think that of the most part, things like neutrality and the desire to have reliability and no original research and civility and things like that… my impression is that they are largely the same everywhere. Things like references are enforced more strongly in some languages, especially in bigger ones, like in English, the rules for citations and references are more strongly enforced. I've heard that is also in German, but I don't know German very well, I've heard German, I learned a little bit of German on Duolingo, but I don't know German very well, so I don't know how well it's enforced, but I've heard that it is enforced. In Hebrew you are also supposed to have sources, but it's just not very strongly enforced, and people kind of... experienced editors that edit a lot, rely more on each other and trust other editors, like sometimes they rely on that more than they rely on external sources, and, personally I don't like that very much. It makes much more sense to rely on external sources and not on other editors, but I guess it has a lot to do with the fact that it's just a small community and people know each other, and people see the same usernames all the time. Unlike in English, where there are many thousands of people and you see different people all the time, so you cannot rely on anybody. That is just my theory, I might be wrong. #00:12:43#

21 IR: But you talk about the quality, is that right? #00:14:57#

22 IE: You can call it quality, I prefer to call it reliability. #00:15:05#

23 IR: Ok, for example, if you compare such a large Wikipedia like the English one with a much smaller like the Hebrew one, would you say that the Hebrew Wikipedia is a part of the English one? #00:15:15#

24 IE: No, no, absolutely not. There is a lot of overlap but they are definitely not the same. It's very easy to check the number of articles that appear in Hebrew and not

appear in English. I think it's around, maybe even half of articles in Hebrew that don't appear in English, I am not sure, but it's very easy to find a number. You can find it on Wikidata. A lot of articles in Hebrew, and the same is true for a lot of languages, they are about local cultures, the culture of the people who write in that language, and it makes a lot of sense. There is even a lot of proper academic research about this, I know at least one person who does this, there might be more. Do you know Marc Miquel maybe? #00:15:36#

25 IR: No. #00:16:37#

26 IE: I can send you the name. He is from Barcelona and he is very active in Catalan Wikipedia community and he wrote several academic papers about this. He did a lot of work about Catalan Wikipedia, but also on many other languages, and he compared the topics, like how much each language is covering the topics that are related to that language. #00:16:38#

27 IR: Sounds a little bit french... #00:17:28#

28 IE: Yes, Catalan culture is a little bit between French and Spanish cultures. I didn't tell you, but I happen to know Catalan language quite well. #00:17:30#

29 IR: Ah. It's not so easy, isn't it? #00:17:42#

30 IE: Hm, it's not so hard. There are a lot of very good learning materials. Like it's a super active language in the international Wikimedia community, as you possibly know, it's the first language that started after English, Catalan and German, on the same day. #00:17:44#

31 IR: Really? I thought that... #00:18:10#

32 IE: January 15th 2001, English started, and then March 15th, Catalan and German started, on the same day. #00:18:12#

33 IR: On the same day? So, if you have the same topic in different Wikipedias, how are the articles connected, apart from the interwiki links. #00:18:32#

34 IE: They are not necessarily connected, they often are, but they are not necessarily connected. Often, but nobody knows how often exactly, articles are translated to other languages, so my team in the Wikimedia Foundation develops Content Translation, which is a tool that helps people to translate articles, and few days ago we celebrated translations of 600.000 articles, using this MediaWiki extension. So, 600.000 articles were translated, but we are absolutely sure there are many many more than 600.000, in all of the languages. People are translating from big languages to small languages, from English, German, Spanish, French, Russian into small languages, and also from small languages to big languages. This is often forgotten, but happens quite a lot. It's difficult to find how many articles exactly, I actually tried to run a research project with somebody from Hebrew university. She made some kind of a methodology to do that, she ran it on the sample, like some things with machine learning, clever methods I don't know much about, she ran it on the sample of the Hebrew Wikipedia, and she found some very estimated, imprecise numbers of which articles are translations and which articles were written from scratch. So, it's difficult to know. Because, there is no proper structured way to know which ones are translations and which are

written by people from scratch. That is just how it is. Of course, there is also the question of how much is it adapted to the culture of the people who speak that language. So, for example, when people translate articles, they may skip some paragraphs that are not relevant for the target language, or they may add paragraphs that are relevant for the people who speak that language. So, for example, if you are in the articles about Paul McCartney, in the English Wikipedia article, you will probably say that he was touring a lot around the world, because English is a global language, but if you write the same article in the Hebrew Wikipedia, it will definitely mention that Paul McCartney played a concert in Israel in 2008, right? This probably won't be mentioned in the English Wikipedia article. #00:18:50#

35 IR: Ok. Don't you think that, if we foster more translations in Wikipedia, that it will change the character of the project, because, in general, it is a project where every community has its own Wikipedia and its own knowledge. And if you only translate from one Wikipedia to the other, from the English Wikipedia, for example, then there is a kind of domination of the English culture in these articles, don't you think? #00:21:45#

36 IE: It's not a problem, if the people do it with the right attitude, it's not a problem, and let me explain why. It definitely changes the character, but not in a way that most people think. The first big encyclopedia that was published in Russian, in many volumes, more than a hundred years ago, it was the Brockhaus encyclopedia, much of which was translated from German. #00:23:03#

37 IR: Which year? #00:23:46#

38 IE: More than a hundred year ago... probably late 19th century... so some of it was written originally in Russian, but most of it was translated from German. And that is ok, because, thanks to this translation from German, a tradition of writing Russian encyclopedias started. And later, in, I think, in the 1930-ies, Russia started producing its own encyclopedias written from scratch. Would it be possible for Russia to start writing their own encyclopedia, if it didn't translate an encyclopedia earlier... I'm not sure, I think it was a good kick-starter for creating the idea that you can have an encyclopedia in this language, so translation is mostly a good thing. #00:23:48#

39 IR: Ok. But, maybe then there is content which will not be visible, for example, I have another case, do you know WikiHow? #00:24:50#

40 IE: Yes. #00:25:07#

41 IR: I talked to an employee, and she said that the main version is the English one and there is a big community of volunteers, and other versions are simply translations from that. And, if you see what kind of content they deal with, it's knowledge about the world in a way that you say... well, how do I do this and that, how do I deal with little problems in my culture. But you don't have the context of this special country and culture, but you have the English ones because of the translation. #00:25:46#

42 IE: Yes, but in the Wikipedia that is ok, because, if you don't prevent people from writing about their own culture, then they can translate useful parts from other

languages, and they can write the other things by themselves. That is ok, that is what we do in Wikipedia. That is also what was done with Russian Brockhaus encyclopedia, they translated a lot from German and they added the necessary Russian parts by themselves, and that worked very well. #00:25:58#

43 IR: And they do this? People do that? #00:26:29#

44 IE: Yeah, that is what they did, that is exactly what they did, if they had to change something, then they changed it, they weren't forced to translate everything word-by-word and not to add anything by themselves. They bought the rights, they translated whatever they needed, and they changed whatever they needed and they added other things that were relevant for Russia and that was fine. #00:26:39#

45 IR: But the reason why translations should be fostered is mainly that its quicker than to write the content originally? #00:27:06#

46 IE: Yes, its quick, it's easy, and … #00:27:19#

47 IR: Ok... the ordering systems in Wikipedia, they are quite different as well, right? Or are there projects to make them corresponding somehow? #00:27:21#

48 IE: Like different, like Wikipedia in different languages? #00:27:45#

49 IR: The category system for example, they differ in the Wikipedias as well. #00:27:51#

50 IE: Yes, that is true and that is a bit painful. I'm not so happy about this. There is no particularly good reason why that is like that. That is just how the software was made back in 2003 or so when categories were first created in MediaWiki. A much bigger problem are templates, the fact that templates are not global, categories should also be global, but the fact that templates are not global, that is a total disaster. That is very bad. #00:27:51#

51 IR: And there are no initiatives to fix that somehow? #00:28:33#

52 IE: Well, there is an initiative which I am leading, I wrote a lot about why this is a problem, and how this should be fixed, I can send you a link. But actually fixing this would be a pretty significant software development project, and I believe it will happen someday, because there is no choice, you have to do this, it doesn't scale. There are templates that do not have to be the same in all languages, that can be unique to some languages, that's fine, but there are templates that really should be the same in all languages, like authority control, and some infoboxes and some citation templates, they should be the same everywhere, but currently they are not, which is really bad, which makes translation much harder than it should be, and it makes sharing knowledge across languages much harder than it should be. It makes starting a Wikipedia in a new language much harder than it should be. Because currently, if you just start a Wikipedia in a new language, you just get a bare installation of MediaWiki with a few extensions, but not templates, and a lot of the functionality in the big Wikipedias is in the templates, like infoboxes and citation and navigation boxes and lots of other things. I saw this in new languages, they spend months if not years setting up the templates, this is just insane, it's a terrible waste of time, so templates really should be global, like images are global. Images have been global in Wikimedia Commons since 2004,

and nobody is complaining about the fact that images are global, nobody is even calling them global images, people just call them images, it's taken for granted, and the same should be done with templates. Some images are local, there are some situations in which certain Wikipedias want to have local files and that's fine. But most images are global in Commons and it's good for everybody. #00:28:38#

53 IR: Where are they stored if they are local? #00:30:58#

54 IE: They are stored on the wiki. A simple example is movie posters, so because of some copyright rules, you cannot upload movie posters to Commons. So, if you see an article about a movie in the English Wikipedia, the image of the poster at the top will be stored on the English Wikipedia, and not on Commons. #00:31:03#

55 IR: Ok. What forms of translation are used currently at Wikipedia? Manually, of course, probably a lot, automatically as well? #00:31:35#

56 IE: So, people can create an article from scratch and write whatever they want in that article, some people will just open two windows on their laptop and look at one article, and just type the translation into the new article, and some people will copy the text of the source article, put it in Google Translate, translate it and paste it from Google Translate into the new article. Hopefully, they will correct the machine translation before they publish it. Sometimes people do it without correcting the machine translation, which is very bad, and these articles will likely be deleted. And there is Content Translation, which my team is developing, which helps you translate between two languages... have you ever seen Content Translation in action? #00:31:50#

57 IR: We also discussed it within our company, there are many customers that want to use multilingual concept within their company, so this extension is very often observed by our technicians. #00:32:53#

58 IE: That is great to hear, you probably have to modify it somehow because the way we developed it is very tied to Wikipedia, but that is good to hear. Lots of people use that, we just celebrated 600.000 articles a few days ago. It helps people just see both articles on the screen at the same time. It allows people to use machine translation for a lot of languages, not for all, but for a lot of languages. It is enabled for all languages, even if there is no machine translation at all for them. If Google translate does not support it, then.... but there is a Wikipedia in that language we allow people to use it anyway, just without machine translation. We very strongly encourage people to correct mistakes machine translation makes, because it makes a lot of mistakes. It automatically adapts links, it automatically adapts images, so you don't have to upload the image again, or type the filename of the image, you just click and it's there. We tried to adapt templates, but because templates are not global, it sometimes works and it sometimes does not work. It is far from perfect; this is one of the reasons I think templates should be global. This is used for a lot of languages... and there is one more thing I really want to mention, even though it's not exactly about Wikipedia content, but it's related nevertheless, there is the translation of the software on Translatewiki. Are you familiar with that? #00:33:22#

59 IR: It is another case I will have to look at. We know Raimond Spekking. #00:35:09#

60 IE: Yes, he is one of the maintainers, he is absolutely great, I absolutely admire what he does. He is amazing. So Translatewiki is not for content, it is for software. However, it's really important to mention this, because software... you know, for the end users who are not involved with writing Wikipedia content, for them, it all blends together, like the menu around the article and the article itself, it's all text, so it's related. Also, as you may or may not know. the MediaWiki extension Translate is also used for translating pages. In theory it could be used for translating Wikipedia articles, in practice, it's not used for that, because Translate extension is very sentence-level, and it adds a lot of markup to the translatable page. It‘s good for technical documentation, like for help pages, manual pages... stuff like that, but it's not very good for Wikipedia articles, because many people who write Wikipedia articles don't want to deal with these tags and all that markup, it just gets in the way, so we don't use it for articles, but it is a thing that should be mentioned. Precisely because of this: for technical documentation it's useful to have the exact same document, word for word in different languages. For encyclopedia articles, there should be more flexibility, precisely because of the reason you said, because different languages may have different needs, different emphasis on different topics, so you can skip some paragraphs, you can add some paragraphs, and translate extension is much more pedantic about having the same, the exact same thing in both languages. #00:35:22#

61 IR: Yes, and I think you also don't have to have as much control over the versions. There are documents, where every edit is really vital and you have to check it in other languages, but... #00:37:32#

62 IE: Exactly, this would be good for Wikipedia articles, if there were many more Wikipedia editors in all the languages, but in practice... if there were many more Wikipedia editors who would very quickly check for all the changes, all the necessary changes and make the updates, than that would possibly scale... but even then, you would still have the problem of... like is there a master article in some language, and culturally that would not work very well... if English is the master language, lots of people wouldn't like it for very good reasons. I wouldn't like it. #00:37:56#

63 IR: As a feedback for your extension, a lot of users would prefer a kind of mixture between Translate and Content Translation, because they want to have this link between these articles. If there is a change in one language, we need to check it in the other, and they only have it in Translate, and in Content Translation, the articles are separate after the translation. #00:38:33#

64 IE: Yes. I know this very well. This is a totally sensible request, and we were thinking about this. When the language team started in Wikimedia Foundation in 2011, from very early on, we thought that we should develop something like the Translate extension but for Wikipedia articles. At first, we thought about adapting the Translate extension for Wikipedia articles, but we just couldn't think of a good way to scale the change management for Wikipedia articles. And then we just decided to give up on this completely, just make Content Translation as a tool for

creating the first version easily and to release this and let people use it and learn from it, and sometime in the future to think about change management. And we really don't regret this decision, because thanks to this we now have 600.000 more articles, and around 3.000 articles are written using it every week, so this is good, and we are thinking, right now, about some kind of a system to update articles, by translation after they are published, we are now working on something like this. It will take us many months to release something, but we are working on this actively right now. #00:39:08#

65 IR: It would help just to have the information that there are changes in the original version, it doesn't have to be so much markup and so on, as you said, but just the message there is a change, or even that there is no change, it's a message that is helpful. #00:40:46#

66 IE: Yes, we are working on something like this, we call it “section translation”, it's a big project, it sounds simple but it's not simple at all. We call it "section translation" because it would allow translating separate sections of articles rather than the whole article and one of the features that will be there is identifying that something changed in the original article and you should update the relevant section. So, it's on the way, but I cannot tell you when will it actually happen, but it's definitely on the way. #00:41:10#

67 IR: And how can I get the news? #00:41:54#

68 IE: That is what we are working on, we are not really sure, we are currently at the architectural stage, we will probably... I can just send you the link to the designs and you can see. I am less involved, I am involved, but not as much as I used to have been in the past in the detailed designs of the current development of Content Translation. I am in the same team, but I'm going to do slightly different things now, so I don't remember it by heart, but I can send you the designs, it's all out in the open, so you can read it. There will probably be some machine learning involved, we are talking about this with the research team. We could in theory just check for changes, but this probably does not scale very well, because there are so many changes all the time in all Wikipedia articles. We are trying to have some smart heuristic guesses of what would be the most relevant changes to actually emphasize to the users, and have them read it, because if we notify them about all the change, it will probably be too much. It's a really difficult question, but that is as much as I can say about it right now. I can send you the links to designs. #00:41:59#

69 IR: By the way, our time is up, it's 12 o'clock, if you have anything else to do, then please tell me... #00:43:27#

70 IE: I do, however, I saw that you have a lot of more questions there, in the list, and lot of them are relevant and I would be happy to answer them, so if you want to have another conversation like this sometime soon, I would be happy to do that, just put something on my calendar. #00:43:38#

[Goodbye]