MONTREAL – IDN Latin Generation Panel Meeting with Integration Panel
Monday, November 4, 2019 – 12:15 to 13:15 EDT
ICANN66 | Montréal, Canada

SARMAD HUSSAIN: Thank you all for joining the meeting here on an update from Latin GP and a discussion between Latin GP and the Integration Panel. How would you want to start, Asmus?

ASMUS FREYTAG: Asmus for the Integration Panel. We had a little side discussion right before the meeting started and we agreed on really heavily focusing on the discussion part because that's the most urgent one. The IP is fully up to date with what the Latin GP has done because we just have a recent proposal. So we thought we might just start with having the Latin GP pose a question because we also have a feedback document we sent. It's also being sent out electronically. This is just a backup copy for those who find it easier to follow it in print. Maybe we should start with having whichever members of the Latin GP have a technical or procedural question with regards to this process and work from there.

BILL JOURIS: I wonder if the IP could help clarify something for me just in looking at the whole subject of variants. In a number of communications, the IP has referred to the "reasonably careful user" with, as I recall, the comment being if you have two domain names that differ by one

character, if the reasonably careful user will notice, then it's not a variant. If the reasonably careful user will not notice, then it is a variant. My question is, what is your vision for what portion of the population falls into that? Ten percent? Five percent? One percent? One-tenth of a percent? What are we aiming for here?

ASMUS FREYTAG: I must give you my apology. I thought I had an example of a reasonably observant user that I brought with, but I seem to have misplaced him or her. It's a construct. So what I think would be more useful, and I'm sorry if that shortcuts your question, but I think it would be more useful if we look at a couple cases that have come up, specific and concrete ones, and see whether we can come to a joint understanding of whether these are variants or not and that by implication that defines the level of distinction that we think might be applied. And if the question is still open at the end, let's pick it up again.

BILL JOURIS: Okay.

SARMAD HUSSAIN: Before I guess we get into the substantive discussion, I was wondering if it was okay for IP members to very briefly describe, summarize their feedback just to get everybody on the same page and then maybe dive deeper into – so not everything is captured perhaps in this document. You've read the proposal. If there's any general feedback and then any

specific points you want to highlight before we get into actual point-by-point discussion, that may be useful.

ASMUS FREYTAG: Sarmad, that sounds like an excellent idea. Very briefly, we have some general observations in our feedback. We have looked in detail at the variants that are introduced because of underlining and proposed by the Latin GP. There's a long list in Section 4.1 on that.

We have also looked at implicitly some other [inaudible] that show up in these sets and in some cases we are questioning now seeing it in the context presented whether certain code points are maybe better left out [but] three of them out of the Latin LGR background their issues seem to be difficult.

There are some other issues we found with variants. Those are listed in Section 5. And we spent some time analyzing the proposed variants for the sharp S which is the only allocatable variant in the Latin GP proposal.

I would almost suggest if we want to take the feedback to start the discussion that we don't go in forward but in reverse order. I think the issue of the sharp S may have had the most unexpected feedback from us in terms of what the Latin GP was maybe thinking. And even though we are potentially agreeable with the starting point, we think there are some other things that need to be added.

Those might be most in need of some explanation, so it might help if somebody from the Latin GP who has seen this could suggest what

aspect of our review we should clarify and we can work from there and then maybe work through other items in this document in reverse order. I think that would be a good thread to follow for this discussion.

So let me ask the Latin GP is there anything about our suggestions how to handle overlapped variants and avoid multiple allocatable variants that you find in need of some discussion? Or do you perfectly understand what we're after and can implement it immediately?

DENNIS TANAKA: In the spirit to get this rolling, on the sharp S the one thing that in particular caught my attention that there was no mention in your feedback and I'm not sure what part of the document you read but about the inconsistency in resolution depending on the browser that the end user might use and you commented more on the linguistic aspect of it. So I'm not sure that was an accident or you intended not to comment on that issue, especially because that was the ask from IP to look at the sharp S in the context of IDNA 2003 compatibility. And primarily our solution to make those [variants] of each other primarily was to solve for that.

I get how to minimize allocatable variants. That's a more fine tuning of the rules as [inaudible] [evaluation] rules, and that's okay because it will not undermine the overall principles these two characters are going to be variants. So if you could just clarify whether that was left out intentionally or it's not [to play] here. Thank you.

ASMUS FREYTAG: Thank you, Dennis. That makes a very clear question. I see three different aspects to the variant with the sharp S. There is a variant proposed of it to the beta, and we will have a chance to hear the Greek Generation Panel's take on that one. I would call that particular proposed variant, well, it would be a good case for the stuff that's just on the edge of what a variant might entail or not because of it's not so much the browser. It's the different fonts that present things differently.

And in that aspect, we do understand the conception of the psycholinguistic realities underlying users of that Latin script that I think the Latin GP has done some great effort at teasing that out. And as I remarked yesterday in the meeting with the CJK Panels, the psycholinguist realities of scripts are different for each script. Users are trained different ways and, therefore, can expect to pay attention to different things.

So that's the one aspect of it. It's mapping to the [inaudible] beta. The other aspect is its alternation in German language and Swiss language orthographies with the double S and its occasional alternation as a fallback with the double S in German as well. That is one aspect, and that led to your design of the one-way allocatable variant which I think is on the face of it not objectionable. It seems reasonably well thought out and seems to match the linguistic reality. That's why we focused on describing that in our document.

But it clashes with a number of other things, and that's the third one. It is an overlapped variant because the S itself has variant mappings. Therefore, any time you map SS to something you have to also map to

all possible strings of variants of S at the same time. Otherwise, your variant label sets are not going to be well behaved. So at first cut, we look at the variant code point sets and make sure they are symmetric and transitive. But in order for blocked variants to work the way we want to, the variant label sets also have to be well behaved.

And we ran into that with some of the [index] scripts where we worked out you might call it a bit of a theory of what is required. A good part of that is documented in RFC 8228. If you haven't read that, you probably should. Whatever we discovered since is documented in LGR-3, Section 6.

So that explains a couple of the steps necessary to make sure that those kinds of variant sets involving sequences are well behaved. I have extracted some of that here in feedback for you in 6.2 on overlapped variants.

DENNIS TANAKA: Asmus?

ASMUS FREYTAG: Yeah?

DENNIS TANAKA: Question, a clarification. When you say well behaved, what does that mean? Does that mean that every variant relationship is in relation to labels that are created there is symmetry and transitivity? Is that what we're looking for or something else?

ASMUS FREYTAG: Well, that's why I pointed you to the RFC and other documents because it's a little involved. So simplified, yes. What you want is if you have a label, you want the label and all its variants – and with some blocked variants the number of blocked variants could be well in the millions – so you want the label and all its variants to form a symmetric and transitive set that is disjoined. That's the most important part. It's disjoined from all other label sets. So each label is a member of only one variant label set.

So that's the first thing. It has to be a member of only one. And the second one is it has to be possible no matter which member of the set you start from, you can compute the same index variant. There's a specification for how that is done. So that requirement of a unique set and a unique index variant is a little weaker than fully symmetric and transitive, but only slightly so. So if you have a fully symmetric and transitive set for the labels, then you're fine.

BILL JOURIS: Just one process question, are you looking for us in our document to go through all of the implications of transitivity? If we have A is a variant of B and B is a variant of C, do we also need to include in the document that A is a variant of C, or will transitivity merely work that out for us? Which way do you need us to document it?

ASMUS FREYTAG: Okay, that's a fair question. The principle we've been following with all other LGRs is that we expect that in the RFC 7940 XML format, you fully specify the symmetric and transitive variant relations. You can use an automated tool to help you fill in the blanks. That Is not a problem as long as you then clean up the output of that tool and present it as the XML.

Specifically, I can mention this right now, if you use a tool that puts comments in there that are something like automatically added for transitivity, we do not want to see those comments in the final published LGR. They have no place there. They are helping you understand the output of a tool. They're not something for the end user.

BILL JOURIS: Okay, thank you.

ASMUS FREYTAG: Now for the specific question of the sharp S and its overlapped thing, it generates quite a bit of extra entries. However, you are lucky that you have only one of them. I mean really you're lucky. We have had some other LGRs where things were not that simple.

So let me ask, does the Latin GP understand our feedback on the sharp S and feel they can act on it, or do we need to explain or add anything more? I see a lot of silent nodding of heads.

DENNIS TANAKA: I think I get it. Like I said, it won't change the overarching relationship, but it's more adding the fine tuning of subtyping such that it affects when we have labels which ones are going to be allocatable as a label and which ones should not be allocatable as a label. For example, when you have this combination two sharp S which are the ones that really need to be allocatable, so it's going to be the label that has the two sharp S and the label that has full mapping to the SS but any other combination should be blocked.

ASMUS FREYTAG: Right. And if you start with a label that has two sharp S and one SS naturally, which can happen in German, then the things that are allocatable would be the one with the maximal number of sharp S and the one with none. So that's how that would work. I have had some offline discussions with some of you on the mechanisms of these mappings. If you have any uncertainty how to specify that in the XML format, please contact us with a quick direct question to clarify something so we don't have to go through the laborious process of doing a whole [draft] just to get a clarification.

We do expect that you explain the general subtyping mechanism in your document, in the Word document, the end user readable document. But you can possibly steal some of the language from our feedback if you want. We don't have copyright on that.

So if that is the end, I was going to move to something else. But let me know if there's still something pending on this issue before we close it. Okay, good. I would suggest we quickly go to – next to 6 we have Section

7 and 5. We can deal with Section 7 later in the context of another code point. Let's go look at Section 5. There are three subsections to it.

If we go in reverse order, there is 5.3. We found that the document has an explanation in the appendix that two code points should be a variant and it explains at great length why that should be and nothing is found in Section 6.5 or the table or the XML. So that's basically a straightforward one. Make up your mind. Is it or isn't it?

BILL JOURIS: I think it is, but in the current draft we haven't gotten around to putting it in the main part of the document.

ASMUS FREYTAG: Okay, so be aware we notice those things. Good. So that clears that one up. I think the position of the IP on this one cannot be final because it hasn't been formally proposed. So I would rate it as some of the members of the IP find it plausible. Leave it at that. We don't know what we finally will say. We may come back saying we can't wrap our mind around it. And that's not to give you grief. That's just the way the process works.

Okay then, we have Section 5.2. There's a possibly mistaken reflexive mapping. For some reason Latin small letter U with hook has been designated an out of repertoire var, which really would remove it from the repertoire effectively. Which is fine for things that could be parts of other scripts or repertoires but not for a Latin character. So we hope

this was just somebody's typo. Or maybe you do want to remove that code point from the repertoire, and that would be fine too.

DENNIS TANAKA: I think it was a typo.

ASMUS FREYTAG: Okay, so it looks like you have to doublecheck your records to be sure. So that's the feedback we gave, so that's simple. And then the rest of Section 5, there are two sets of variant sets that involve I think Vietnamese usage that we in our review have flagged as we are potentially not on board with seeing those as variants. One of them has the O with hook and a sigma identified as variants. That's something we can discuss also with the Greek GP when we have the meeting. And then the other one is the breve with the circumflex and grave. We found that a little hard to swallow.

So I think these are kinds of things where we basically can say we have problems with it and then I think what we would expect from you is to either say, yeah, fine. Maybe those were in the middle of the gray zone. We can color them a little darker and rule them out. Or you can come back and say, no, we have really strong reasons for those and give us a bit of reasoning in writing for that. So what we are saying is if you just repeat it in your next draft and don't have a particular reasoning why these ones in particular are needed, we may just not accept.

Okay, so then we can go and I think let's go look at in the Section 4, at the bottom of it, not at 4.17 but at 4.16. Because that is one that has a

similarity to Section 7. The review flagged a concern with the Latin small letter gamma. Not all reviewers were happy with seeing that mapped. I guess the reason is in typical fonts it was coded as a separate character because it freezes a particular form of the Greek gamma that isn't found in all Greek fonts and that's the reason for it to exist. But it does not necessarily make – so adding that or not adding that doesn't complicate any sets. It's just an observation that we felt it was not necessarily [done]. This is one where if you really feel like you want it, you give us a bit more of an argument on it.

DENNIS TANAKA: So that specific item I think came out on the URL on the [line] analysis and some of the GP members thought that there was a visual similarity or confusability issue.

ASMUS FREYTAG: So our general position on the underlining issue is that we are following you in your analysis of detached items and we're having more difficulty with connected descenders. And I think that reflects here. But, for instance, if there was a well-known font combination where the actual Greek gamma and the Latin gamma looked the same, that would be good to know.

BILL JOURIS: I don't think we actually considered the Latin gamma and the Greek gamma per se. We looked at the Latin Y and the Greek gamma….

ASMUS FREYTAG: Then you looked at the Y versus the….

BILL JOURIS: Then we looked at the Y versus the Latin gamma, and that's where we got there.

ASMUS FREYTAG: Okay, so basically our position can be summarized we're a little doubtful on the connection of Y [into] Latin gamma. That's the one we're a little doubtful on. If you can find a strong reason to map it against the Greek gamma, let us know what that reason in and maybe we can work from there. So that's our feedback.

Then the last one in that set is the Latin small letter Y with dot below. While we generally are saying, we say that in the opening section of Section 4, that we follow your analysis in having disconnected dots treated as if they are made indistinguishable by underlining we think the Y with the dot is really special because the dot is placed to the right of the descender which makes it confusable with a trailing period. So if you have a label that ends with that, it will look like a separator in the domain name at that point. Or if you have it in the middle, it can possibly look like there's an oddly formatted label separator. And we are really concerned with that, as useful as it is in that particular orthography, it is possibly too risky for the root zone based on that.

It's clashing with the sequence [Y period], so that's why we are having it listed here struck out and for you to consider and perhaps you come to an agreement that, yes, it's in the hyper-conservative realm of the root, we may have to take that out. And it's similar to O2BC, the modifier letter apostrophe that we had to take out because it clashes with punctuation.

So this is one where the part that seems to clash with the punctuation is built into the glyph. And in Section 7, we have two more of them. Which is kind of stupid we only mention them now because, of course, we've been aware that they are presented that way since the early days of Unicode when we got the feedback because we had listed them with actual carons in the code charts and we got feedback that we're using the wrong glyphs. So all of us kind of knew that these things had that comma shape.

We have tried to show how if you take some French words, you can create reasonable facsimiles of these French words using the D with caron instead of a D and an apostrophe. That would, like the N with the comma above for some of the African languages would basically be one where we'd be more happy to not have that in the root for that reason. Difficulty, we understand they're very central to the Czech writing system. But on the other hand, the root is the root. So that's our feedback. Yes?

BILL JOURIS: Yeah, can I go back to the Y with the dot below? Is there a font which we didn't look at where the dot below is actually not in the underline but

actually up higher? Because the fonts that we looked at, the underline went through the dot which would seem to eliminate the concern. If there's a font we missed, then we can look at that. But we didn't see anywhere it was even possible to see that dot.

ASMUS FREYTAG: Well, there are many times that domain name labels are represented without underline. So you're correct in the sense that if it is underlined, you may not see the dot. But if it isn't underlined, you will see the dot and then you have the confusion with the punctuation mark.

BILL JOURIS: Okay.

ASMUS FREYTAG: Some of these concerns are kind of orthogonal and if one applies, then we should pay attention to.

BILL JOURIS: So that's basically a case to take the Y with the dot below out of the repertoire. That's not a variant issue; that's a repertoire issue.

ASMUS FREYTAG: Then that would be a repertoire issue, year.

BILL JOURIS: Okay.

ASMUS FREYTAG: And as much as we don't like to – I mean, with alphabeting languages, taking out a whole character is always a grave thing. But for the root, the amount of mischief you can create is just maybe overriding.

Since we have discussed descenders, I am going to skip – yeah? You have a question?

UNIDENTIFIED FEMALE: I have a question about Topic 7. It seems the same that we should exclude from the repertoire letter D with caron and letter L with caron for the same reason.

ASMUS FREYTAG: For the same reason. And if you want to write this up, you may mention also – we may have done this already in MSR – but there is an exclusion of the letter N with a comma above in an African language. That's also excluded. No matter where it was excluded. You should probably cite that in the discussion so people understand that it's a systematic thing.

UNIDENTIFIED FEMALE: Yeah, okay. Thank you.

ASMUS FREYTAG: I was going to suggest since we started discussing the descenders, we skip to the beginning of Section 4, yeah right about here. Where it says cedilla [organic]. So a half page up, please. Good.

So as part of looking at the underlining logic we felt that applying it to disconnected descenders made a lot of sense because if you have something that is basically as high as the width of a single line, then it is really easily covered by that single line. But we were less persuaded by the reasoning to extend that to connected descenders like the examples we have here. The Latin letter thorn, [ing], and the gamma.

I will tell you that with the Cyrillic letter, the IP is not unanimous yet in our evaluation whether the Cyrillic letter and the U with dot below are not close enough to be treated as not readily distinguishable. So we have an open mind on that third item. It is not really marked here. But like you, we were a little bit in a time pressure to get our feedback out. So I'm telling you this for the record. The third item here is one that we may be persuaded that, yes, it is a variant of U with dot below.

However, that may introduce some transitive variants, so we need to see whether the cost of that – the more transitive variants are introduced, the more you want to look at the original premise that caused the transitivity is strong. And if it somewhat marginal, you may decide to split. And I'm sure you've had discussions like that before. Bill?

BILL JOURIS: Yes, these are in fact the ones that occasion might question about the reasonably careful user because we certainly didn't have consensus either. In the case of the cedilla, there are maybe seven Latin letters that can include that particular diacritic. And on four of them, we concluded they were variants and on three we did not. And all of those were 4-3

votes. So the question, then becomes how sharp-eyed do you assume the users are that you are willing to hang them out to dry?

ASMUS FREYTAG: Yeah. So it's if you had had 9 people on your committee, you would have a 5-4 vote on those. Which however is perfectly where I want you to be in this discussion. It's an indicator that that particular distinction is in the gray zone. So once you have these very split votes, that means that you can go and look at other effects. So if it introduces a lot of transitive sets, then maybe it's not a low-hanging fruit so you might as well not make it a variant. Basically that.

So for the cedilla, we looked at this and it's here in Section 4. It's unfortunately so long there. So there are a couple of places. The general thing on the cedilla that we're seeing is that a cedilla and comma below should be variants of each other. We think that is appropriate. And the dot below should be a variant of the base character.

And then here on this thing you can see we looked at the case. Some of the code points that were introduced because of the fact that the orthography for Marshallese is not fully stable and there are important works Marshallese that are published with a dot below where a cedilla below should exist and vice versa. So here you have a suggestion of possibly taking out some more items from the repertoire.

I want to have Pitinan.

PITINAN KOOARMORNPATANA: We have questions online. This is from Meikal: "Did IP consider the visual data presented in the appendix for this? To clarify, there should be graphics in the appendix in which proposed variants are compared across different fonts as evidence."

ASMUS FREYTAG: Yes, we looked at all those pieces that actually printed correctly. There's a whole appendix of supposed data that didn't show up. So that would be one – I mentioned this to some people offline – you have this big spreadsheet that contains all sorts of interesting data. Maybe that can be published as separate files first. The actual spreadsheet can be published as a file. The Chinese LGR has 20 separate files that they publish, so there is no way you can be exceptional in what you're doing. I would suggest you do actually do two. That you publish the XML as XML and publish a printout into PDF of it so that it freezes the fonts that you've used on your master system for the evaluation. But publish it also as spreadsheets so people can actually look at the data easily. That would be my recommendation.

There are some of the appendices that might work as separate files, but many of the appendices are perfectly fine to keep in the document as is. It's always more convenient if things are in one place. But if it's just a long table or long dataset with not a lot of text, that would be always a candidate for outsourcing, farming out.

BILL JOURIS:	While we're on that section, there's a comment at the bottom of Page 2 there about the comma below only occurring in Romanian and the Romanians have no problem making the distinction. And that seems to be somewhat in conflict with the point in the general remarks about characters may occur, especially deliberately from bad actors, in things done with other languages.
	Which way are we looking to go here? Because you said something similar to us about Vietnamese as I recall, about these are all confusable but the Vietnamese can tell the difference so it doesn't matter. I think you're right in the general remarks here that, yes, it does matter because they're not going to be the only ones using it. But which way would you prefer us to go?
ASMUS FREYTAG:	So this looks like a bit of – I apologize - it looks like a bit of sloppy editing by the IP. You're absolutely correct. Characters can be presented out of context. Take the Vietnamese case. If you have two Vietnamese heavily accented [forms] that look like each other to people who are not Vietnamese, then we would be fine saying if the Vietnamese can tell them apart, we're okay as long as don't look like some character used in a third language.
	Now the comma below looks like a cedilla, and the cedilla is clearly used in a third language. So here that argument does not apply. In that case, whoever drafted this piece of text had a mental short circuit here. Very good for you to note that. And, please, if we make nonsense

arguments, you don't follow them, and then tell us where we're wrong and we'll get a better LGR as the process.

BILL JOURIS: I think we appreciate that you were trying to rush this out and so it didn't get the editing that it normally would. I'm sure you found cases in ours. We already noted that one XML case where something was done a little more quickly than might have been ideal.

ASMUS FREYTAG: Right. But I mean, we're luckily far away from having this go to public comment tomorrow so we can all catch these things and work on them. As much as possible we're attempting not to present you with moving goal posts, even though sometimes it looks that way. But the issue is each time we get a proposal, we look at it with somewhat fresh eyes in its overall context and we read the explanations that exist and things. And in the meantime, we may have had experiences reviewing some other LGRs that brought certain issues to our attention, and that can lead to the fact that we're suddenly aware of an issue that we weren't before. So if the sharp S had come at us three years ago, we would have probably let it go through the way you proposed it. We weren't understanding of the issues that we raised today. We had to learn those with [Malayalam] and [Tamil] and others.

BILL JOURIS: I think we understand the goal posts move, and in some cases I at least feel they need to move. So I don't think that's a flaw. I think that's just

an acknowledgement that you weren't omniscient when you started out. And as we go on, I think we'll keep finding things where, oh, we hadn't had that brought to our attention so we need to rethink this. I think that's inevitable in the process.

ASMUS FREYTAG: As long as we can manage with some result at the end. We had a discussion yesterday in the CJK Panel meeting, and it was all captured in a single number: 200. Mark is smiling because the exchange was how many users do we need to poll to get the perfect answer in terms of distinguishability, and is it a failure if we only have 199? Then Mark said, "200?" And then the generation panel promised to 201 to be on the safe side. But that's the kind of perfection that we're not trying to get to that level.

BILL JOURIS: We're only going to get to that level with the public comment where I think we're all clear that we will get lots of comments from people who are native speakers of languages that none of us speak and probably have never heard of until we went through to look for the repertoire. And that's inevitably going to happen.

ASMUS FREYTAG: It's either that or you discover it after you publish. So we have so far avoided most of those issues. We have one small issue with one LGR that's published, so we're working on that.

Good. Let's see where we are. We now have left in Section 4 a lot of detail. I invite everybody to scroll through. I'm going to pick up Section 4.1.11 because it presents a new issue that we haven't discussed. It's the one that has blue in it. It's easy to find. There we go.

Your text indicates that you're aware of the issue but haven't really finalized it. The issue is that we have a number of other scripts that have a circle shape that is confusable with the O. It is particularly pressing because of the fact that there is already a delegated TLD .ooo. So it's not hypothetical. It's an actual one that exists, and we need to make sure that we remove spoofing.

Now the way it works process wise, we've asked the other LGRs not to bother putting that into theirs. We will wait for your to put these ones in, and that kicks it into the integration process and they get imposed on all the other scripts. That is fine. We just didn't want them to have to work against what to them felt like an unknown target because the Latin script didn't exist yet in terms of draft. So don't feel odd that you're suddenly the one [asserting] something about existing scripts. There has been agreement that we were going to do that that way.

BILL JOURIS: Just so you know, we did share with several of these groups what we were doing, so they're not flying totally blind here.

ASMUS FREYTAG: Okay. So I think the three that we have here are the ones that came out of the analysis. We have written down that analysis somewhere. It's

either in an MSR or in the LGR overview document, I forgot which one, for the full table. What we're driving, we're kind of cajoling everybody to all agree on that one set and put that in.

You would include the O with macron below. We will have no problem with that. We are having a little bit more problem with the cedilla. If push comes to shove and at the end you are explaining to us why the shape O really makes a cedilla so hard to distinguish with underlining compared to other things, maybe you can clarify that a little bit more for us. Or you can agree with us to take that one out.

I would suggest that these are the ones that kind of were high on my list to communicate with you and doublecheck that we are on the same page as to what the feedback means and where it's coming from.

I know of one issue that isn't here because it can't be because it wasn't written down in the data tables. Our review is to a large degree based on the data tables. So if you have some thoughts that aren't reflected in the data tables, we don't know that. Meikal, we were discussing which one. There was one Armenian pair that you were wondering whether you could map the two of them, right?

MEIKAL MUMIN: Yes. I think what we did was say the Armenian group is no longer around so we can't talk to them about it. So we'll map the one that's basically a homoglyph and ignore the one that differs by one pixel so nobody but an Armenian would notice. But perhaps in our meeting tomorrow we can talk to them about that now that they're available.

ASMUS FREYTAG: Yeah, we need to bring that up because it won't be in any of the materials that I've prepared because my materials are based on existing data. I'm hoping that you guys will help make sure we don't forget asking them.

I agree with you that this is a case where literally it's the one pixel. My imagination would be we would ask the Armenians how many words they have that are minimal pairs that differ in only that character. And from the names of the character I expect there will be very few, but I haven't done any data analysis or anything to know. If you were in the one percent level, I think it would be feasible to ask them to accept that as a restriction on their set for the greater good, so to speak.

We've done similar things in Ethiopic where Amharic has to have some variants that affect up to one percent of the labels in other of the languages. We felt that was an appropriate level of strictness. So if it's in that level, it's fine. If it happens to be a very common alternation and there are dozens of pairs of words that do that, then we would maybe say, fine, we don't want to necessarily cripple Armenia.

So that can come out in discussion. The good part is that Armenian status at the moment is deferred, not published. So it in some ways was kept deferred to allow for precisely this kind of stuff. So this is all above board. The process is ideally they would actively import your proposed variants. But if they don't for some reason, we are as IP allowed to impose it cross scripts based on if you put it into the pool, the tool generates a superset. And if any GP didn't actively import it, then it gets

imposed. But it doesn't prevent from doublechecking with the other GP that anything imposed shouldn't be allocatable on their end, for instance, or to discuss with the effects are too devastating on the use of labels.

So that was the one that isn't listed here. So where are we in time, and are there any more questions? There are five minutes left, I hear. Does anybody have any questions on any of the other sets in Section 4 or any of the other items in the document or anything else they ever wanted to know from the IP?

DENNIS TANAKA: I don't have any more questions on the document that you sent us, so thank you for that. Anyone else? No? But we are about to embark to undertake visual analysis within the Latin script. So I think it comes down to get us back to what is the guidance there. What is the overall principle? I mean, the careful and observant user, who is and how do we really [distill] what we even look at? Some will propose the base character with the base [character] and the acute when from a visual standpoint we can clearly see they are different. But some would argue that there's some similarity. So I think if you can help us navigate those waters, that would be very helpful.

ASMUS FREYTAG: Okay, so let me try to give you a little bit of input. First, I need to remark that one thing the IP never does is preapprove. So if I say something looks good here, that's all tentative. It will depend on how it's written

down, how it's argued, and how it's presented to us. And then all five members will have to be unanimous before you get an actual approval.

Having said that, let's look at the F hook with F. I find that an interesting case given the psycholinguistics of the Latin script because you can expect Latin users to be so familiar with very small set of basic characters that there has been quite an acceptable range of deviations in the way fonts approach the basic structure of a character. We have the two forms of A, two forms of G, and clearly we have two forms of F. That users rather readily can be expected to accept them. Side-by-side they see the difference immediately. Maybe not with an A and a G, but they see it with an F and the F hook. I would say people see that there's a difference.

But if not presented side-by-side, the F with the hook, I think your analysis is probably good that it is very plausible for a user to see that at the point and they're saying, well, this is just an unusual font. Maybe even not an unusual font. It's just a different font from the rest of my text.

The same incidentally applies to – that's my personal take on it – the I within I dot, dotless I and I. I have come across so many texts where people are using for some effect I guess fonts that don't have a dot on the I, and nobody will object. In book titles and other places, people don't object to that. They just read it, and it looks somehow fresh.

So those psycholinguistic effects I think are reasonable to consider. What should stop you from going too far down that rabbit hole is a different principle that fights against that one. And that is you want to

not put something into the variant mechanism that couldn't reasonably be appealed in an application process. So if you have a thing where you have reasoned yourself several steps away from direct visual equivalence, the further away you go if it wasn't baked in the LGR, if it was in some other process document, people would possibly appeal that.

As much as possible, we want to be restrictive in the root. We want to take obvious cases off the table by outlawing them in the LGR. But we don't want to be in a position that we run into stuff that would try to get people to overturn the fact that the LGR is designed to not be appealable. We don't want to create pressure to say we need to throw the whole thing out because they made this big mistake there. So that should be the break that you apply that you don't go too far.

So you have a couple examples here on this one. The F with F hook, as I said, seems to be plausible. We haven't reviewed it formally because it wasn't in the list. The OE and OE and AE and AE, well, we have predominantly speakers on the panel that find those distinctions rather easy. So that's where we are at that one.

We've discussed some of the stacking concerns. If the stacking only applies to within a single language group, it can be ignored because those users are definitely trained to look for those distinctions. It's my feeling that, I know you said you're starting to look at this, but it's my feeling that there isn't much scope for additional discoveries of really solid, oh, this should be equivalent cases. Surprise us, but we're not

expecting a lot. So if you come up with 50 variant sets where you're all saying these must be variants, we would possibly push back a little bit.

BILL JOURIS: The problem, as I see it, for all of us and I suspect for the five of you is we know too much. There are a couple of dozen different diacritics that get used. And after three years, even those of us who aren't trained linguists know what all of them are, know what they look like, know how they differ. But for the average user, he will know the handful of diacritics that are used in his native language. Maybe he knows a couple of other languages, not counting English which is useless for this purpose.

ASMUS FREYTAG: It's definitely useless, yeah.

BILL JOURIS: Yes. So he'll know the diacritics for a few more and maybe he's up to 8 or 9 out of that 20-some-odd. If he's in Europe, he probably has picked up on a few more that occur in languages around him. The circle over the A in Swedish or whatever. But he's still at under half. For the rest of them, he doesn't even know they exist.

So while it's physically possible to see the difference between, say, a dot and an acute if you have only seen an acute on A, E, O, and You, and you've never seen it on an I, having the dot above the I replaced with an acute is something on the order of the difference between a serif and a

sanserif font. That looks decorative. It's not that you can't see it. It's that it will never occur to you that such a thing is possible.

But the problem we have is trying to get our heads around what does the bulk of the population even know to look at. That was part of the same who's the reasonably careful user. Even a reasonably careful user if he's never seen such a thing, won't realize that it's something that he ought to be looking for.

ASMUS FREYTAG: So you basically answered your own question. If you construct some sample labels where it's a word in English or some other language and you substitute in that word a different accent, does that look like that word still or does it look off enough that somebody might be cautious clicking on it? That's basically what we're trying to do.

And you're not the last line of defense. The variant system is intended for things that people unquestionably take as substitutes. And there is supposed to be a string similarity review of which we know very little that should take care of some other possible cases. So we have to make sure that we don't try to solve more than is our remit. So obvious cases, in your face.

I would say for me the line is basically I with dot, I without dot is amazingly easy to confuse people with. For the acute and the grave on the I, there's at least something there. So you see something. The missing dot is missing, so it's hard to see that it's missing when you're not expecting it to be missing. So that is to me further down the gray

scale. And once you start saying any kind of accent is confusable for any other kind of accent, I think you're in the string confusable review territory. That's basically what I would think off the cuff.

SARMAD HUSSAIN: Just doing a time check here. We're five minutes over time. We also have a comment or a question from Meikal. Do you guys want to take that?

ASMUS FREYTAG: Yes, please.

SARMAD HUSSAIN: Okay, so let me read that. "Does IP potentially agree with the Latin GP's view that a significant majority of fonts rather than a majority or key fonts in which two code points have nearly visual identical glyphs constitutes a risk high enough for the LGR to establish a variant relationship between such code points?"

ASMUS FREYTAG: That was slightly difficult to parse, so let me try to rephrase it so I understand it. So the IP is definitely on record saying that there are a couple of key fonts, and the key fonts that we would be concerned with are the ones that are actually used in the key browsers in address bars and stuff. So if any of those sets of fonts have significant visual overlap, they should be considered.

And we are, I think, coming around to saying that in a few cases where the Latin script throws a letter form that is of the overall variability for that base letter and made the frozen form a special letter in another language that in some of those cases, like possibly with F hook, the case can be made that those are variants too.

But there are all sorts of lovely fonts that do all sorts of crazy things with space shapes, and I wouldn't necessarily say that we would expect to look at the general universe of text fonts. We're looking at user interface fonts specifically and things that possibly URLs are commonly represented in, and those should be the primary ones to look at.

UNIDENTIFIED MALE: I'm sorry. I have to excuse myself. I have a presentation to do in five minutes elsewhere.

UNIDENTIFIED MALE: Well, good luck. I don't know whether that remark I just gave was clarifying at all or obfuscating further. It's a judgment call, but we would focus on things that are focused on the typical situations where users are exposed to labels and have to act on a visual representation of the label to recognize it to choose to follow it or do something with it.

So basically, the only time we're open to considering other types of fonts is if they're indicative of a really generic shape variability for a base character. Like with the F with hook or the sharp S with descender, which by looking around you can say, yes, there are basically two forms

[inaudible]. And while the UI fonts all have one, there's enough ordinary life exposures that users can have that in their mental state.

But I would limit the second such observation to really clear-cut cases. And if you identify too few of them, we're better off than if you go completely wild and give us hundreds of them and we basically say, yeah, they're medium-gray but on the wrong side of the line.

So I see now again lots of head nodding all around the room, so I think we have reached the conclusion of this. Thank you, Meikal, for dialing in and joining the conversation. Thank the members of the Latin Generation Panel for their hard work. It's definitely turning into a Ph.D. thesis, that proposal. We thought Latin might be a little bit difficult, but you guys – anyway, so looking forward I'm hoping that we will see rapid progress so that we can not miss the timeline for the next upcoming LGR. Because our process is going to get fraught with much more difficulties in the future, so the sooner we can get this done the better.

SARMAD HUSSAIN: Yeah, so let's [inaudible] that before we disperse. I think the timeline we're looking at is early January to start the LGR-4 integration process. So it would be great to get input from Latin GP sooner than later, I guess, if you want to meet that target.

It takes about five to six months to conclude the whole process, taking a GP proposal to public comment and then integrating public comments back into the proposal. Finalizing the proposal, and then the IP doing the final review, and then integrating that. And then a public

comment on the integrated [inaudible] LGR. So the whole thing takes about six months.

We are hoping to do a new rules on LGR release by end of June, so we'll start the process in early January. So that's something to think about, see if that's an achievable goal. Thanks.

MIRJANA TASIC: I have to add something to this. I don't expect that we can finish it until June next year. Sorry.

ASMUS FREYTAG: Okay, that puts an interesting spin on it. We may have to discuss what we will do because the Latin is a little bit of an anchor [tenant] in whatever LGR [inaudible] because it drags with it the Greek, Armenian, and Cyrillic immediately. So in that context, that makes it even more urgent to [warn you] that you've identified a number of obvious in script variants in Latin already. You're not getting too many bonus points for going too deeply into the nonobvious cases. There is a perfection level that's the enemy of the good at that point. So I want to warn that as well.

SARMAD HUSSAIN: All right, so on that positive note, let's adjourn this session unless there are any last comments around the floor. No? All right, then great. Thank you all for joining, and we hope this session was useful. Thank you.

[END OF TRANSCRIPTION]