NIKO PARTANEN / HZSK

FEW REMAINING ISSUES WITH CURRENT WORKFLOWS

Notes

TO SUMMARIZE, I have selected two distinct topics which I think have to be addressed somehow. However, large portion of this handout describes our basic corpus workflow. Reason of this is to make rest of the conversation more sensible, as I'm afraid some ideas seem weird or unnecessary if dropped without further context. The topics I address specifically are:

- Reuse of corpus raw data – how files go together
- Citing the individual resources in the corpus

I must mention already now that I do not have any finished solutions for any of these issues. But in many cases we know where the problem is.

I also want to emphasize that there is nothing new here. All infrastructure around the things described basically exists, but it is more matter of adopting practices around these tools. And of course some things just do not work as they should right now.

Introduction

THIS PRESENTATION is a small discussion about few aspects in the current state of our documentary linguistic data management workflow. Actually I will not focus very much into data management, but more to those aspects of it which are relevant for the later use and reuse of that data. There is this endless talk about standards and best practices, but I'm not so certain how useful this discussion always is. More important than which standard is used is to be consistent with the chosen model. At least to me it is becoming obvious that if I want to work with some new dataset, among the first things to do is **to transform it**. I want to read it into some other structure which is compatible with some other dataset I may be using. Thereby it is quite trivial what is the original structure, as long as it is consistent, easy to understand and well documented. The last two are never really true, but on the other hand if consistency is good it is pretty easy to work onward. Similarly best practices are best evaluated later with comparing the ready datasets, the results of that work, and descriptions of chosen workflows – although in many cases this is not necessarily possible.

I'm also not sure if all my observations make any sense, it is very possible that I view some topics from a strange bubble and the others are doing things very differently, but I assume at least some ideas I

discuss here have wider relevancy. For example, I don't really know what kind of corpora other people have been building and how they are generally used. I have my impressions, but I can't assume these are necessarily very accurate. Please keep this in mind while, as I make here some claims while I speak and please correct me when I'm wrong.

For the use of our datasets I see there are few key issues which connect to the data management.

- The use of individual id's for utterances
- Enough metadata to distinguish the most salient features
- Explicit relationships for different files in the corpus
- Clear system of corpus versions so that the data can be reused
- All files should adhere into same structure

It is kind of funny that the definitions of word **corpus** contain all kind of fancy explanations, but they usually do not mention that each file in the corpus has to look the same. This is something so trivial that it doesn't even need to be mentioned, but in many ways it is something that we are very easily failing with.

The language documentation outcomes are in many ways relatively similar with one another and also other corpora, although it is also clear that there are big differences. But when we talk about differences I'm not sure whether we have actually really compared different corpora and concluded that they are very different. I even encountered recently an idea that *language documentation outputs are not actual corpora*, which of course would raise the question what are they. I would assume any language documentation dataset contains the following types of information:

- Transcribed utterances which form longer texts
- Transcribed utterances and words which are parts of elicitation sessions
- Translations for at least some of the utterances
- Recordings which are not transcribed or annotated
- Information about the recording place and time
- Information about the speakers (sex, age, birthplace)

Maybe all this information is not always present, naturally there may be few speakers for who it is not obvious where they are born, for example. And with old recordings it can be vague where they were done by who, but that is another topic altogether. But in principle I cannot comprehend how it could be possible to collect any kind of data without having information like this. So if we reduce the demands into bare minimums like these the datasets are actually quite comparable. At least one can start to answer questions like:

How many speakers there are?, **What is the speakers sex or age distribution?**, **What is the ratio of transcribed and untranscribed data?**, **How many tokens and types there are?**. Naturally there are also often glosses.

For transforming corpora into different formats lots of things hang with these two or three pieces of information:

- Session id
- Speaker id (and if possible speaker role)

Everything else is extra, but without these being super-consistent it is quite impossible to understand what is going on.

Our Komi ELAN corpus

Our Komi corpus contains transcriptions of conversation / semi-structured interviews[1]. Their average length is around 45 minutes and we try to transcribe the complete events. We are still experimenting with this, but the idea we have had has been that if one session is defined as uninterrupted stream of recording then that's how we should deal with it whenever possible.

[1] This is a problematic term in itself – it is an easy way to describe any not particularly structured conversation, but it is not so often discussed what in hell we ask in these interviews

This has the side-effect that many individual units we often want to describe, for example to have in metadata, such as **story**, **elicitation** or **song** are entirely meaningless as categorisations of individual sessions, but can be used to refer only to small parts of it. In this situation one has to device some kind of elements that span across the recording portions and mark where different units start and end. This is quite necessary since complete transcribed recordings can be quite difficult and messy to navigate.

Topics in our corpus are mainly biographical and could be loosely termed as ethnographical[2]. We have selected these topics because we can assume that people are often willing and comfortable to discuss them. We avoid sensitive and political topics in order to ensure the reusability of the data. Usually the profession or hobbies of individual speaker leads us to focus to the relevant themes, which also brings thematical variation to the interviews. Biographical data is readily usable in the sense that Iźva Komi have expanded to their current territories in not too distant past, the process has been active just a few generations ago, so family biographies often reveal lots of details about how this has unfolded. Also discussions about different ethnicities in the family can be very informative, and this is a natural point to discuss the individual's language skills and history of language learning.

[2] Although it is obvious that our work doesn't have very much to do with ethnography as a scientific field. Old people speaking about how they used to live is not ethnography.

The sessions are recorded, usually with lapel microphones, and we use one or two video camera to capture the video. The video has

also been a place of experimenting for us, for example, in the last fieldwork at Kola Peninsula Michael was shooting more closer views with mainly the speaker in the frame, which is bit different from our earlier recordings where we tried to have the most of the speakers in the view. Problem was the the video composition often wasn't that nice when everyone was in the picture. One could argue that one needs to see everyone in order to study the gestures, but in the same vein one could say that to study gestures we need really exact video for all speakers from correct angles – generic video that direction may not be sufficient for these uses in the end anyway.

Typical setup for our sessions has been that there are several audio and video files associated with one recording. Video file also contains its own audio track(s), which also can be useful in some situations[3]. The problem with lapel microphones is that a lonely lapel mic gives excellent sound for the speaker on who the microphone is, but the other speakers may be relatively quiet. The only solution is to mix together different tracks to end up with something that is pleasant to listen. In this situation it should be very well documented how different raw media files relate to those files which are used in transcription and with which the ELAN file is associated. This data is stored in the export files of those software used in file processing, in our case PluralEyes 3 and Final Cut Pro X, but it is not trivial question whether this data can realistically be digged from those files in case one really needs to recover it. Naturally this also implies that these exported XML files are stored and archived with the corpus itself[4].

[3] In some cases the surround sound can be wanted

[4] This is a good question to ask around: **Does anyone archive these files at the moment?**

Topic one: Post-processing workflows

First task is to gather all related audio and video files and to synchronize these. This can be done with many different tools, ones I discuss here are just one option. However, this is a very good practice to do very fast, as this way one sees instantly if some file is missing and if something is wrong with the tools. At the moment I would actually recommend really sitting down and listening to the files carefully, and having the video on big screen and really thinking and commenting it instantly on the same evening after it has been shot. We need more practices like these so that our work actually can improve and evolve.

I will now show how the basic data synchronization workflow looks like. First one synchronizes the files in Plural Eyes, I use one of our recent recordings as an example (Blokland et al. 2015a). What has been done here is that all files related to the session have been put into one folder, and that has been thrown into Plural Eyes. Most of the time this is very straightforward and the files align nicely. The audio waveform is used to to do the alignation, so everything the file

needs is some audio. It doesn't need to be good audio. So in principle one could have in camera very bad microphone, I often use my DSLR which makes really horrible sound, but that is totally ok to get the video aligned.

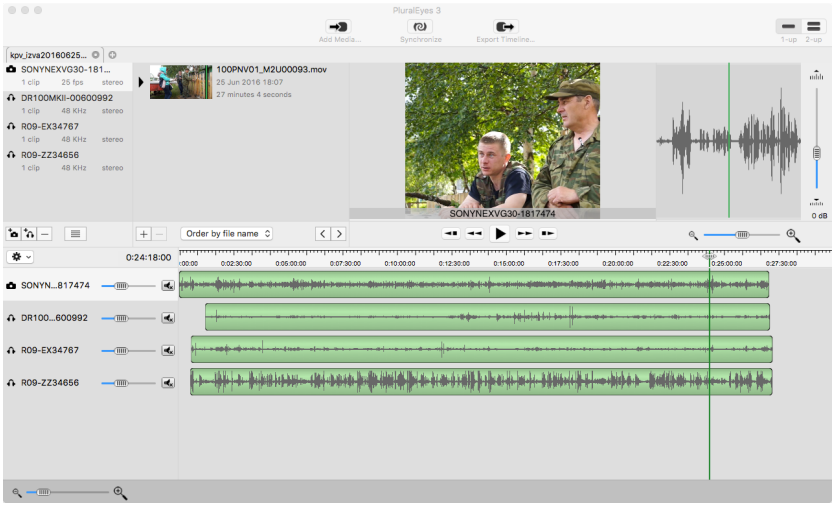

Figure 1: Plural Eyes after succesfull synchronisation

From Plural Eyes it is possible to make export to different programs, for example to Final Cut Pro or Adobe Premiere. As far as I know Premiere can do something like this on its own as well. In theory I think so can also Final Cut Pro X, but my results were not that good when I tried it the last time. Maybe I didn't know what I was doing! Anyway, maybe the step above can be avoided. Clicking the exported file opens it in Final Cut Pro.

Our current idea has been that for the version which is transcribed we would not do too much cutting. Instead the transcription would be done to that stretch of raw data what there is. If there is a portion without video then there can be just blank screen. The problem here is that we also want to make other video versions which look nice, are enjoyable to watch and sensible by their content. These are cut from the same raw data as the ELAN files, but how do we keep information about this exact relationship which exists between the files?

As far as I understand it, in this specific case the answer lies in content of Final Cut Pro XML export files, or in the Premiere equivalents. It has to be possible to parse these files and reconstruct how which file has been created from which source files, which again can be matched with the file used with the ELAN file. This isn't simple, of course, but *this is too complicated* is not an acceptable argument in this topic. The Final Cut Pro X XML contains elements like these:

```
<spine lane="-1" offset="1791/25s">
```

```
        <clip duration="46637591/30000s"
                    format="r1580D7D6-0A6E-4B77-BB31-470BB0DE5428"
                    name="160625_0709.wav" offset="0s" start="0/48000s">
          <audio duration="1554600/1000s"
                    offset="0/48000s"
                    ref="r0E551A9F-5213-4373-8C92-0AA982728CEA"
                    role="dialogue" start="0/48000s"/>
        </clip>
</spine>
```

This somehow contains information about the filename, and the first offset must be related to when it starts in relation to the other files. Somewhere else it is indicated whether the audio has been muted or volume turned up for this particular track and so on. Of course working with this kind of files is real pain in the ass, I so much don't want to even start to write some script that manages these files and extracts information from there, but it can be done. In principle it could be even written in session metadata for the individual files. Saying basically: **Hey, this file is used in this session with this kind of settings**. So in principle if there are some movie makers, for example, who want to use our video in their own work, it should be quite straightforward[5] to pick up this kind of information from their export XML, be it from FCPX or Premiere, and use that to associate the annotations from our ELAN files with the video which is another kind of constellation of those same files. If these programs can take XML like that and reconstruct this, then we can also do that:

[5] although not simple!

Figure 2: Same project in FCPX

This is important also for later use of files. Usually the audio associated with the ELAN file is mixed from different sources. It must

be instantly evident whether for this session there is a lapel mic track for speakers X and Y, in case someone wants to do, for example, phonetic analysis of speech of those individuals. At the moment we are not keeping adequately track about which microphone was with which speaker, on the other hand also because this is immediately evident when one listens to the recordings. It is easy to know that yes, this is the lapel mic of that speaker because it sounds like a lapel mic. However, the computer has no way of knowing this, and we probably would need to store this information in more machine readable format.

One advantage of multiple video is that it can be used to document what has been done. For example, with our recent data we have been experimenting with following setup (Blokland et al. 2016):

Figure 3: Two videos side by side

In principle one could also have video aligned like this in ELAN. And underlying XML which has the information about the video alignment could be used to signal where are the segments where multiple video or audio is available. Naturally this all gets more interesting when one starts to record events which are more spontaneous and are not just interviews/elicitation sessions, but we haven't got that point yet. We have some situations where we are close to this and getting to good and new direction.

Topic two: Citing the data

I've seen there have been gazillions of workshops about data citation. I'm not exactly sure what people have discussed there, but in my opinion one of the most acute questions is practical. **How do I cite the resources I use while I'm researching something and actively writing?** There are different cases which demand different approaches, but often we are told to cite corpus with some conventions

which contain the corpus name, year and authors. In reality this cannot be so simple, as I think in many corpora different files have quite different authors, they are published originally in different publications and recordings have been through quite many hands before they end up to be combined with the current project. In some cases it may be sufficient to cite them as belonging to the project as such, but in reality this will probably lead to many problems as we usually probably have to mention the original publication and those authors, although we would had ourselves worked with the same data as well.

And of course one can have in bibliography an entry like:

```
@incollection{PSDP,
    Author = {Joshua Wilbur},
    Booktitle = {Endangered Languages Archive (ELAR)},
    Keywords = {Database;Saamic linguistics,Text collection},
    Location = {London},
    Publisher = {Hans Rausing Endangered Languages Project},
    Title = {Pite Saami: Documenting the language and culture},
    Url = {http://elar.soas.ac.uk/deposit/0053},
    Year = {2008+},
    Bdsk-Url-1 = {http://www.mpi.nl/DOBES/},
    Bdsk-Url-2 = {http://www.hrelp.org/archive/},
    Bdsk-Url-3 = {http://elar.soas.ac.uk/deposit/wilbur2009pitesaami}}
```

Or:

```
@incollection{KSDP,
    Author = {Rie{\ss}ler, Michael and Scheller, Elisabeth and Kotcheva, Kristina...},
    Booksubtitle = {{DoBeS} archive},
    Booktitle = {Endangered languages},
    Hyphenation = {american},
    Keywords = {Kola Saami,Corpus-sjd;Corpus-sjt;Corpus-sms;Database;Saamic linguistics...},
    Location = {Nijmegen},
    Note = {[Digital archive]},
    Publisher = {Max Planck Institute for Psycholinguistics},
    Title = {{Kola S{\'a}mi Documentation Project (KSDP)}},
    Url = {http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI363060%23},
    Year = {2005+},
    Bdsk-Url-1 = {http://corpus1.mpi.nl/ds/imdi%5C_browser?openpath=MPI363060%23},
    Bdsk-Url-2 = {http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI363060%23}}
```

And this is a very good and totally works. One can cite them very easily: Josh does great corpora, see for example (Wilbur 2008+).
Under the hood this is something like this in LaTeX or markdown:

```
\cite{PSDP} or [@PSDP]
```

Citation info like this is nowadays associated with most of the corpora. There are major problems, for example, that these are not counted as publications and do not count in academic crap-rankings of who publishes how much, which unfortunately is a problem one has to care about. But there are also other issues. One is that this is not really how we use the data! I don't want to cite just the corpus, I want to cite individual items in it! At least individual sessions and also individual utterances within the sessions. I can see situations where someone wants to cite individual tokens or other annotations in the corpus, but I think we are smart enough to find the right location as long as we have access to the individual utterances. This means that each utterance has to have some kind of id.

One aspect in this is that the id's are only valid for specific corpus versions. I may go tomorrow around my files and spot some typos, delete some utterances which do not contain anything, and this is probably enough to change the utterance id ordering of the file. Thereby one has to be able to refer to exact version one was using while doing the research. In this point it is often stated that it is necessary the corpus is *frozen* (Kübler and Zinsmeister 2015). However, a spoken language corpus can never be frozen since the interpretation of the spoken language is such a subjective and easily changing issue. Of course one can pretend that the transcription is final, but if we have any larger texts it is certain that one has to change quite a bit every now and then, especially when someone is actively using the corpus.

I'm not talking today about storing the corpus in GitHub, although this is a very good convention. Each change in corpus creates a distinct **commit**, and each of them has an id. This could be one ultimate way to refer to specific corpus versions, but the world is probably not ready for this at the moment. But I think in the world of language documentation we do not employ at all the concept of corpus versions at the moment, and this is a major disadvantage if one wants to use these datasets in a manner which is reproducible and thereby reliable.

I have now been citing several recording sessions along the way, and one way of pointing to the individual sessions can be seen there. However, what about citing individual items? In principle one could use very similar mechanism there. So something like `\cite{IKDP}` could bring up this: (Blokland et al. 2014), but one could also refer to something like `\cite{kpv_izva20150703-01-b}`, which would bring up this: (Blokland et al. 2015b). This is a session citation. This is obvious as it confirms to our session naming scheme, which contains ISO-code, dialect tag, date, order number and tag for which portion of the recording we are talking about. Sometimes a longer session has been split into several units, although we don't really want to do this, but as we have seen things aren't that perfect most of the time.

Now, this session contains an ELAN file. This ELAN file contains transcriptions, of course, it looks like this to make this clearer:

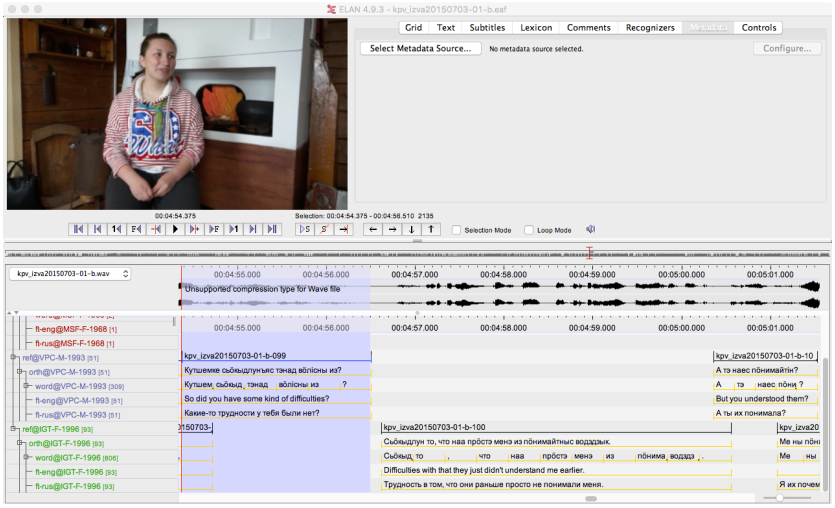

Figure 4: ELAN file

For the sake of example, let's pay attention to wonderful pronoun /*naa*/ in example kpv_izva20150703-01-b-100. In standard Komi we would have /*naja*/, whereas in Iźva dialect which we have here one normally finds /*nia*/. What is the exact distribution of /*naa*/? Everyone would love to know, but I think no one has looked into this yet. However, we can use this as an example of citing individual corpus item.

> In Iźva Komi /*naa*/ is not typical, but it occurs everywhere, for example in speech of Salehard Komi, as exemplified in (Blokland et al. 2015c).

What I want to emphasize here is that for a reference to the segment I want the output look a bit like one below. Now I'm using there segment time start and end, but of course one could also use the reference id. The question is bit complicated, because what you really want is **a link that refers to that item where ever it is**:

> Blokland et. al. 2015. "Session: kpv_izva20150703-01-b. Interview with Irina Terentyeva." Segment time start: 296623 ms, time end: 300565 ms. Recorded 7.3.2015 in Ižma, Komi Republic, Russia. Resource created in Iźva Komi Documentation Project, funded by Kone Foundation in 2014—2016. Archived in `http://hdl.handle.net/00000/0000-0000-0XX0-0`.

Ideally the link in reference would actually be linkable and lead, for example, to that annotation in Trova. I can't test that now as the LAT seems to be under maintenance, and anyway I would not

have my current Komi files there because the upload system has been so difficult, but I exemplify this with Finnish corpus. In LAT this is adviced to cite as: *"Suomen kielen näytteitä - Samples of Spoken Finnish [online speech corpus], version 1.0. - Helsinki : Institute for the Languages of Finland, 2014. [accessed dd.mm.yyyy]. Available at: http://urn.fi/urn:nbn:fi:lb-1001100134."*, and this is also all good and fine. But if we pick up examples from there it is often quite important to be able to refer to individual examples from the file.

> Finnish dialect spoken around Mikkeli, Savo, is the most beautiful variant of any Finnic language. For reference, see this narrative about skinning a squirrel (Suomen kielen näytteitä - Samples of Spoken Finnish [online speech corpus] 2014).

Now if you click that link, which I also put here as in PDF it goes to the bottom (in HTML it should be in the side), one should get that segment open.

The important idea here is of course that these citations should be accessible through a bibliography which is generated from the annotations and the metadata. It should be possible to refer to the items like this without thinking twice. It gets more complicated of course, because some materials have only been digitalized in the project, some have been transliterated, some have been published before and reused etc, etc.

Appendix: Some other pressing issues

MOST ACUTE issues at the moment are, in my opinion, related to following questions.

How to store research related annotations as corpus enriching annotations? I mean that now we often make a search, annotate the result and continue onward, but shouldn't those annotations be better stored within the corpus itself so that others could also benefit from them (plus reproduce the findings)? The main issue is that necessarily the corpus version gets out of sync with the version one is externally working with. This is in some sense connected to the question how to manage the glossing/annotating work done outside the main format where the corpus is stored and worked with

I think the use of video has to be theoretized more in language documentation, especially as it complicates the workflows as shown above. It is somehow obvious that video adds value, but **how**, **why**, **in what ways?**. Something I've thought about is that apparently with ethnographic films there have been something called **written companions** (Heider 2006), which means longer written document

which explains the events in the film. Maybe we could employ something like that, explaining what happens in a recording, what are the topics covered, tying that into literature as well. For example, when people are talking about how they herded reindeer with the Nenetses and which languages they spoke in that context, one could associate that recording segment with the citations to different sources where this topic is discussed or mentioned. In my opinion this would add very much value to what we are doing. Of course this is enormous work and it can be difficult to convince people that this is worth the effort, but I'm moving to this direction myself anyway as I think it will be interesting.

Our current workflow neglects photographs entirely. Adobe Lightroom could be worth trying out, as the newest version has the face recognition which is certainly useful for us, but I haven't been happy with any solution by now. Images would need to be classified in relation to the sessions, but also tagged for people, events, items, practices, topics etc.

Can we somehow connect our sessions better to the real time? So instead of saying that a recording was done in 1.1.2015 and lasted around an hour, couldn't we say it was done between 2016-07-18 17:43:06 CEST and 2016-07-18 18:51:34 CEST? This would have additional benefit that all photographs and similar content could also be easily time aligned with the recording, as they have their own timestamps.[6]

[6] Of course this isn't so simple, as there is no way to synchronize different devices…

I think that now once we have worked several years with Komi and from the technical viewpoint lots of things are working well, it is more acute than ever to start to focus a bit into content. I would assume that not everyone agrees with this, but for me it has always been extremely interesting what these people actually tell. And when I spend time going through some Ludian data or Kven data or Russian dialect data, it occurs again and again that there are thematically and topically connected items all over the place, and there must be some value in linking all this data together in one way or another.

Blokland, Rogier, Vassili Chuprov, Marina Fedina, Niko Partanen, and Michael Rießler. 2014. "Iźva Komi Documentation Project. Funded by Kone Foundation in 2014-2016. Corpus Archived Somewhere." `http://hdl.handle.net/00000/0000-0000-0XX0-0`.

———. 2015a. "Interview with Aleksandr and Vassili Artiev." Recorded 25.6.2016 in Krasnošelye, Murmanskaja Oblast, Russia. Resource created in Iźva Komi Documentation Project, funded by Kone Foundation in 2014—2016. Archived somewhere. `http://hdl.handle.net/00000/0000-0000-0XX0-0`.

———. 2015b. "Interview with Irina Terentyeva." Recorded 7.3.2015 in Ižma, Komi Republic, Russia. Resource created in Iźva Komi Doc-

umentation Project, funded by Kone Foundation in 2014—2016. Archived somewhere. http://hdl.handle.net/00000/0000-0000-0XX0-0.

———. 2015c. "Interview with Irina Terentyeva." Segment time start: 296623 ms, time end: 300565 ms. Recorded 7.3.2015 in Ižma, Komi Republic, Russia. Resource created in Iźva Komi Documentation Project, funded by Kone Foundation in 2014—2016. Archived somewhere. http://hdl.handle.net/00000/0000-0000-0XX0-0.

———. 2016. "Interview with Nadežda Hatanzyei." Recorded 23.6.2016 in Krasnošelye, Murmanskaja Oblast, Russia. Resource created in Iźva Komi Documentation Project, funded by Kone Foundation in 2014—2016. Archived somewhere. http://hdl.handle.net/00000/0000-0000-0XX0-0.

Heider, Karl G. 2006. *Ethnographic Film. Revised Edition*. University of Texas Press.

Kübler, Sandra, and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.

Suomen kielen näytteitä - Samples of Spoken Finnish [online speech corpus]. 2014. "Helsinki : Institute for the Languages of Finland, 2014. [Accessed 20.07.2016]." Session: SKN10a_Mikkeli. Segment time start: 3618237 ms, segment duration: 39096 ms. https://lat.csc.fi/ds/annex/runLoader?handle=hdl:11113/00-0000-0000-0000-1DDE-2&time=3618237&duration=39096&viewType=timeline.

Wilbur, Joshua. 2008+. "Pite Saami: Documenting the Language and Culture." In *Endangered Languages Archive (ELAR)*. London: Hans Rausing Endangered Languages Project. http://elar.soas.ac.uk/deposit/0053.