Sharing of Data as it relates to Human Subjects Issues and Data Management Plans

Natasha Warner
University of Arizona

I. Introduction: The scope of the question

We are here to discuss how to protect human subjects from whom data is gathered, while also sharing data as widely as possible among researchers, and archiving it in a stable way to prevent data loss and make data accessible to others. The first aspect, protecting human subjects' rights and privacy, has very different implications depending what type of corpus data one is collecting, and from what group of speakers one collects it. Much linguistic corpus data is practically no-risk: if speakers recruited from university classes hold a casual conversation about non-sensitive subjects, such as what they had for dinner, plans for the weekend, how their family members are doing, or how their classes are going, and names are not mentioned, they may not be concerned with whether the recordings are kept private.

However, if speakers begin discussing sensitive topics, or if an interviewer leads them to sensitive topics (e.g. illegal drug use, personal or medical information), the same speakers may be more concerned about privacy. Furthermore, some speakers are by nature at-risk: recording gay and lesbian speakers who are not out talking about sexual orientation would require considerable privacy safeguards, for example.

Finally, some entire language communities have strong objections to publicly sharing data, such as some Native American tribes that do not want their language to be shared with people from outside the tribe.

In this presentation, I will talk primarily about low-risk and no-risk situations, where speakers have few or no concerns about how their data is shared. Situations that truly do put speakers at risk are more complicated, highly varied, and perhaps more deserving of the attention Human Subjects Protection offices devote.

II. Conflicting goals and motivations

Several parties are involved in the question of whether a corpus should be archived and shared: the speakers, the researcher, the Human Subjects office (IRB), funding agencies, and archiving organizations. Several of these groups differ in their goals and motivations regarding how data is handled.

Researchers often have an ethic of making knowledge publicly available. As scientists, we usually want to make information freely available, and share it, in order to maximize the benefit of the work and further human understanding of language. This motivation is very basic to the undertaking of science in general: scientific investigation leads to greater understanding of how the world works, and for that to be effective, the knowledge that is gained through science cannot be kept private.

However, this is not researchers' only concern, and researchers are not selfless. Most researchers also wish to publish their findings and insights. They need to do this before others publish similar results, or their own work will not be publishable. We are all familiar with the pressures of "publish or perish"—our careers, and hence our salaries and house payments, depend partly on publishing. Linguists experience less danger of being "scooped" on a result than scientists in fields such as Micro and Cell Biology, but linguists may still not feel a strong motivation to share their raw data publicly, for other researchers to work with, when they have put the effort in to collect that raw data and have not yet published everything they could from it. If the researcher works in industry rather than academia, or is developing a commercial product using the data, so that the data serves applied rather than basic research, the researcher and his/her employer may have a strong motivation not to share data.

Another factor is that not all of us are impeccably careful with filenaming conventions or with documenting how data is organized, so that the raw data files may make sense to the person who collected them, but not be immediately interpretable for others. Making one's data publicly available requires improving one's level of organization and documentation, which takes time.

Human Subjects Protection Offices (and their associated IRBs) have almost an opposite motivation: to keep data as private as possible, so that there could not possibly be any risk to the speakers. Since federal regulations are often unclear, and Human Subjects staff are not trained in the fields whose research they supervise, many Human Subjects Offices err on the side of caution and make non-sensitive data much more private than it would need to be to legitimately protect the speakers. One factor that contributes to this problem is that Human Subjects Offices have to get their information about a proposed study from the researcher (the PI), but PIs are clearly not a neutral party, and instead are highly motivated to obtain permission so that they can proceed with their research. Thus, Human Subjects staff may not trust PIs to think of every possible risk to speakers. One can see the logic of concluding that data that is kept completely private, or better yet deleted, cannot possibly pose a danger to anyone, so it would be safest when in doubt to force researchers to keep everything private. The fact that this means other researchers will have to record more speakers more times, subjecting them to whatever the imagined potential danger is, is not an immediate concern, since it will be handled whenever those future Human Subjects proposals are submitted. This leads to the situation we have at some universities, where completely non-sensitive data containing no names, consisting of recordings of subjects reading a list of isolated words, are required to be destroyed after a set amount of time because voices are recorded. Ambiguity in federal regulations has also led to sometimes nonsensical assumptions, such as that all human voices are inherently completely identifiable if recorded. It is hard to imagine that someone will break into my office, steal my computer, listen to the recordings of speakers reading a list of phonotactically manipulated nonsense words, recognize their voices out of the entire population of speakers of English, publicize the fact that the speakers participated in the experiment, and somehow do something bad to the speakers this way. It is also hard to imagine that we need to build in protection against this scenario. However, it is not surprising that Human Subjects' Offices' primary motivation is to protect privacy, and thus prevent sharing of data.

Funding agencies have a different motivation: if they invest money in having a researcher collect a corpus of data, they may want that data to be shared as broadly as possible, so that other researchers can answer other questions with it, leading to greater impact from the same amount of funding. This is at least likely for basic research funding, such as from NSF. It is perhaps a less likely motivation for Defense funding.

Speakers' motivations with regard to their recorded data vary greatly, and are of a different sort. If speakers are recruited from an introductory linguistics course, and receive extra credit for participating in an experiment during which they are recorded, their motivation is usually to complete the recording and get the extra credit. They understand they are being recorded, and even if spontaneous conversation is recorded, they usually avoid discussing anything sensitive. Such subjects have minimal concerns about what happens with the data after they finish, especially if researchers only play short bits of recordings publicly or on the web. While we might technically have permission to play a speaker's entire conversation in class or during a talk, including a long discourse in which the speaker fights with her mother about the color of her wedding flowers, we would be unlikely to do this. It is interesting to consider whether knowing that that could happen would make speakers uncomfortable during recording. If we post audio corpus data of spontaneous conversations at an archiving site such as OLAC, LDC or Talkbank, or even directly on our own web pages for public download, are we also making an uncomfortable situation possible?

As mentioned above, if speakers discuss sensitive topics (e.g. sexual orientation, labor unions, opinions about the competence of the local tribal leadership or about one's boss, etc.), they probably have greater motivation to insure the privacy of such recordings. If speakers belong to a community that restricts access to the language, or that wishes to control how community members are used for research (e.g. some Native American tribes), the speaker is likely to have a strong motivation toward privacy. If the speaker belongs to an oppressed minority group in a country where members of that minority group are routinely imprisoned or abused, the speaker would have an extremely strong motivation to keep recordings of themselves speaking the minority language private. However, this is not the norm in linguistic corpus research.

Thus, we see that for most non-sensitive linguistic research, the motivations of the researcher, the funding agency, and the Human Subjects Office are likely to conflict, while the speakers themselves may not have any concerns about the data. Below, I will discuss common outcomes of this situation.

III. Possible outcomes

At many universities, it is perfectly possible to get permission to share one's audio corpus data with other researchers, or even to archive it with an organization that will make it available. At my own university, the Human Subjects Office staff often has the initial inclination to keep everything completely private. However, as long as a researcher describes clearly in the proposal how they plan to share the data, and the consent form explains this sharing of data clearly to the subjects, it is usually completely possible to obtain permission for data sharing. Because Human Subjects Offices so often ask what seem to be picky and unreasonable questions, and because standard forms sometimes ask questions like "when will you destroy the data?," many researchers conclude that they cannot get permission for much of anything, or that they will be required to destroy data.

In fact, "N/A" or "the data will not be destroyed because it poses no risk to subjects, and if it had to be collected again, this would impose a time burden on additional subjects" may be completely acceptable answers. Researchers, especially new ones such as students, seem to get so frightened of Human Subjects review that they restrict themselves from doing many kinds of allowable research before the Human Subjects Office ever has a chance to approve the work. Senior researchers sometimes do the same thing out of frustration.

In fact, it is often possible to do fairly risky research with human subjects as long as the subjects understand the risk completely and consent to it in writing. The same applies to sharing non-risky data: as long as subjects clearly consent to it, it will be fine at my university and many others. Some universities, however, do seem to impose very strict privacy protection with no leeway for no-risk studies, no matter how the researcher explains the situation. This is partly an unfortunate result of unclearness in federal regulations, and probably also simply a result of differences in bureaucracy at various institutions. It is surprising what a large influence the overall approach of the Director of the Human Subjects Office can have on what research is possible, let alone how long it takes to get permission for it.

Thus, three common outcomes of the conflict between motivations of the researcher and the Human Subjects Office are 1) researchers learn how to obtain permission, and thus share and archive raw data with little further difficulty, 2) researchers do not attempt to share data because they assume the Human Subjects Office won't allow it, and 3) Human Subjects Offices do forbid data sharing, and it does not happen.

The NSF's recent change to requiring a Data Management Plan with all submissions has added an interesting new impact to these outcomes. While requirements for the DMP are not yet clear or widely known, the NSF is encouraging greater sharing of data with other researchers through the DMP. It is basically required that the DMP include a promise to share raw data with other interested and qualified researchers (not necessarily with the public). Researchers may specify that they will make the data available a particular number of years after collecting it, or at the end of the funding period, or after initial publication, but they are largely required to include a promise to share it, unless there are compelling reasons to keep the data private. (Another exception would be if the researcher is using existing data, with permission, that does not belong to the researcher.) The DMP may include a specific plan for how and where raw data will be archived following best practices for archiving, or it may simply specify that interested researchers can contact the PI to ask for the data.

Since the DMP requirement has gone into effect just recently, it remains to be seen what effects it will have. However, this is likely to push against the stricter IRBs' requirements of destruction of data. If NSF declines grants where the DMP specifies that all data will be destroyed or not shared, and suggests that researchers go back to their IRB and attempt to get permission to keep and share non-risk data, researchers may have more success when armed with the NSF's panel summary than on their own. Because the NSF is one of the most common funding agencies for linguistic corpus research, the DMP requirement is likely to have an impact. For no-risk data, this may lessen the number of projects that fall into outcomes 1 and 2 above.

IV. The Future and the ANPRM for the Common Rule

The U.S. Federal Government is currently considering a massive re-working of the Human Subjects regulations, with especially strong impacts on low- and no-risk social sciences research. Many of us have been involved with submitting feedback on the ANPRM for the Common Rule in the last few months. ….
As it currently stands in the proposal, the revision to the Common Rule (the current basis of all federal human subjects protection guidelines) would probably improve the Human Subjects permission situation for low- and no-risk linguistic research in many ways, but might make it much more difficult in one way. The proposal includes:

- No-risk research would require only a one-page form notifying the Human Subjects Office that the PI is going to do the research, and the office would not have to approve it before the PI starts the research. (Effectively no review.)
- The PI would get to determine themselves whether their research is no-risk. This has obvious potential dangers, but hopefully universities would find good ways to manage them.
- Consent would be much simpler and less legalistic, possibly just using oral consent for no-risk research.
- Collaborative projects across multiple U.S. universities would only require approval at one university. This has obvious positive implications for sharing data when the researchers who wish to share it are all part of the initial project.
- The one possibly very bad part: imposition of HIPAA privacy requirements for all human subjects data, of all sorts, collected by anyone affiliated with a U.S. university. [[a few reasons why this could be very bad]]

However, we do not yet know what the outcome of the ANPRM will be. Over 1000 comments were submitted on it, from large organizations and from individuals, and many of them object to the HIPAA requirement. Many also praise the proposed changes for no-risk research, although some express concerns with various provisions for various research situations. The working out of what changes to make to the Common Rule, based on the ANPRM and the comments it received, is likely to take a few years to be settled and implemented.

If the changes in the ANPRM, minus the standardized HIPAA requirement, go through, we should see much greater freedom for sharing of non-sensitive corpus data among researchers. Of course, Human Subjects Offices of various universities are likely to implement any regulation differently. Variability in implementation of the rules seems inevitable. If the HIPAA requirement is kept, it may be fairly difficult to make data public, especially as how to implement it is worked out.

V. Conclusions

In conclusion, sharing of non-sensitive speech data is probably more possible now than many researchers realize. The DMP requirement of NSF is encouraging greater data sharing, and researchers may be able to obtain greater permission from their Human Subjects Offices. Finally, possibilities for data sharing, as well as everything else about human subjects permission, may change drastically with the ANPRM for the Common Rule, but the outcomes of this cannot yet be predicted.