Communication Across Great Gulfs of Time, Distance and Difference: Lessons from The Rosetta Disk. L. A. Welcher Ph.D., The Long Now Foundation laura@longnow.org.

Overall Goal and Objective: To better anticipate and understand alien communication systems by developing a better understanding of human communication systems.

In attempting to understand how intelligent life anywhere communicates, it is worth understanding as best we can how intelligent terrestrial life communicates. Not just human beings, of course, because many different terrestrial species communicate, and use a variety of modalities to do so. But given that we modern human beings and our descendants would be primary participants in an extraterrestrial communication event (as message senders or receivers, creators or decipherers) we should certainly try to understand how human communication might work in this context, and how it is likely to be similar to, or different from, alien communication.

My experience provides a few perspectives on this potential enterprise. First, my training is in theoretical and documentary linguistics, and as such I have been a student of human communication for the past ~30 years. Secondly, I practiced this training in the context of the mass global linguistic extinction we are witnessing today, working with some of the most critically endangered languages in the world. And thirdly, I've had the unique opportunity as a linguist to work at The Long Now Foundation [1] on developing The Rosetta Disk [2], a physical archive of all human language, designed to last and be readable for 10,000 years.

The creation of any archive is a communicative act, and its content is a kind of message intended for future generations: "We thought you should have this information so we saved it for you." When that archive is meant to communicate not just to the near future, but over millennia to the very distant future, and the very medium of its message (language itself) is drastically changing, if not disappearing, then the act of archiving as communication with future humans (if you take the effort seriously) becomes an exercise akin to trying to talk with aliens. So I take this as my point of departure and share with you a few of the lessons I've learned so far from this experience.

For a bit of background, The Rosetta Disk takes its inspiration from the original Rosetta Stone used to decipher ancient Egyptian Hieroglyphs. The Disk has ~14,000 pages of parallel-translation documents written in thousands of languages spoken around the world today. The pages are microscopically formed in a thin sheet of nickel, and are readable with optical magnification of roughly 1000x. Around the circumference of the disk, spiraling inward, text in 8 different languages and scripts reads "Languages of the World: This is an Archive of Human Languages." The Disks are a testament to human linguistic diversity at the beginning of 21st century (little of which is likely to exist a few hundred years from now). They are also intended to provide a means of deciphering whatever information we leave to the distant future, in the form of our human language. To date, we have created around 20 of these disks, and they reside in private and public archives around the world.

Not all Rosetta Disks are terrestrial archives. In 02004 we were invited to include a Rosetta Disk on the similarly-named Rosetta ESA mission, which would be the first craft to rendezvous and land on the surface of a comet. Not wanting to miss such an opportunity, we provided what we could at the time, which was one of our first Rosetta Disk prototypes. Its content was 7,000 pages of parallel language documentation, inscribed within a ~2x2 inch square on the surface of the nickel disk. And there it remains today, on the surface of Comet 67P, a testament to human intelligence, bearing its message to unknown recipients in an unknown future.

I'm often asked the very reasonable question, for whom are we creating The Rosetta Disk? Perhaps people figure that anyone "out there" enough to actually create such an object—if sane—must be apocalyptically-minded. Or, perhaps people in our society commonly assume a catastrophic end to human civilization in the near future. Who would be left to discover and decipher such an archive? They are surprised to hear that I'm designing the Rosetta Disk for people thousands of years in the future...our descendants, and that in ways that matter I expect they will be like us.

Is this a safe assumption? If humans are around millennia hence, how will they be like us? Will they have similar bodies? Hands like ours to hold the Disk? Eyes like ours to examine it? Could they manufacture a basic microscope? Then if they can access what is written on the disk, will they be able to decipher it? Will they think like us? Will they use languages anything like ours? Will they understand writing that symbolically represents spoken aspects of those languages? If we look back at the last 10,000 years of human history and development we might expect this to be the case, but given the rapid pace of technological development and our interaction with it, this might not be the trajectory of the next 10,000 years.

If these future humans have languages are like ours (at least in production, structure and organization) we

can make some some educated guesses about them. First, it may not be obvious from casual inspection that they are related to the languages on the Disk and the difference will only be magnified by time. All natural, spoken, human languages change over time (languages that don't change are obsolescent—this is symptomatic, rather than causative of language extinction). Witness only the differences between classical languages like Latin and Sanskrit and their modern descendants, or the difference between Modern English and the Old English of Beowulf—and this with the passing of just a few hundred years.

Secondly, there are over 7,000 human languages spoken in the world today, and if you include dialects of all of those languages, many beyond that. Human language is variable down to the individual speaker. The reasons for this variety in language have to do with the differentiation and change noted above, as well as changes that take place due to contact between speakers of different varieties of language, and the persistent human need to use linguistic variation as a marker of difference and group identity. We shouldn't expect this to change any time soon. Even if we all become speakers of Mandarin in the future, there will likely be other languages in use, and variations in any language use that will turn into greater differentiation—and new varieties of language—in the future.

These are important lessons for both the sending and receiving of interstellar messages. We should expect communication systems might be variable on both ends, with changes in systems taking place over time (potentially considerable change as the time span increases), as well as variability among a single species at any given point in time. But I'd say there is little else we can count on, even with respect to our own species.

Another lesson I've taken from the design of The Rosetta Disk is the important role of message encoding—a consideration for any long-term / long-range communicator. To illustrate, say one wanted to create a long-term archive of the message "the quick brown fox jumps over the lazy dog" using DNA (which is, in fact, being explored as a possible long-term storage medium). First, one would encode this text message in binary to make it digital (glossing over all of the challenges and conventions of binary encoding of the world's writing systems), and then translate the digital 1's and 0's into the nucleotides of DNA. This is actually how information like text, computer programs, and images have been stored in DNA to date.

Then the successful decoding of our message has several challenges, not the least of which is someone discovering that a message exists, and coming up with a means to read the DNA to access it. If our future message-reader hasn't given up altogether by this point, the next task is to figure out that the encoded message is actually binary, that the binary encodes another set of symbols (text), how it encodes them, and that the set of symbols is a representation of spoken human language.

If we go by modern human practice, we might expect that any interstellar message we receive or discover will be encoded, and possibly with multiple layers of different types of encoding. Some of the encodings may be conventional rather than rule-governed. Also there may be no instruction manual. This of course is not best practice, and we should strive with any messages we create to be as transparent as possible about encoding, and to keep layers of encoding as minimal and predictable and reverse-engineerable as possible.

Is The Rosetta Disk better in this regard? It gets around some of these encoding problems by avoiding the digital mode and the additional encoding required by the medium of DNA. It represents text visually, as an analog message. One "reads" the contents of The Rosetta Disk like one would a book. Nevertheless, it still has the challenge of mapping visual symbols (text) to spoken language. And spoken language is, of course, another encoding—that of human experience.

The basic encoding challenge presented by The Rosetta Disk is that of human language itself as a medium of communication. If that language is no longer a lived, spoken language, what are the chances of getting any real meaning out of it? If you are lucky, as was the case with the original Rosetta Stone, there will be lots of other examples of recorded texts, and you can start your task of assembling dictionaries and grammars to attempt to decipher them. This situation is likely to be the exception rather than the rule. But in this way you might learn about people and events from a time and place far distant from your own. It would be a far different challenge to decipher an intentional message left for you from the distant past, and without a speaker to consult, you'd be hard pressed to be absolutely sure of your translation and its meaning.

I'll end these brief remarks not with an answer, but with a question. How do we communicate meaning? This is the problem we are left with even if we solve all of the other challenges of mode, medium and message in long-term communication. In fact I see this as the fundamental challenge of any "long" communication event—across time, across distance, and across potentially vast gulfs of difference.

Returning to the question I'm often asked about The Rosetta Disk—for whom are we creating it? The only thing I expect I can count on about those future humans that might encounter it is that they, like us, are symbol-producers. Languages are wonderfully rich, complex, and organized systems of signs. Language in

turn can itself be symbolic, and humans often wield language in this way. The medium is also the message. I wouldn't expect communication with any intelligent being to be any different. If we can solve this problem of communicating with our future selves, we are likely to be in a better position to decipher the meaning in any messages we discover or receive from other worlds.

Additional Information:

(A) This paper addresses the question: "How does intelligent life communicate?"

(B) Natural language processing (NLP) is an active area of research and application in linguistics and the field has made good progress with speech recognition, synthesis, analysis and translation of a few major world languages. Still, NLP technologies and machine learning approaches intended to interact with them, are fundamentally dependent on our knowledge and understanding of human language, and how it functions socially, culturally, and communicatively.

Major world languages are the best-studied languages. They are also the best documented, and have the greatest amounts of naturalistic data available for computational use. A premise of this paper is that if we are to understand alien communication systems, we should have a good understanding of how our own work. To develop our knowledge of what a possible human language is and how it functions, we ideally need a large amount of naturalistic linguistic data for *all* extant languages, as well as descriptive products such as lexicons, grammars, and analyzed texts. Yet few of the nearly 7,000 languages spoken on earth are well documented, or have anywhere near enough recorded data on which to base NLP or machine learning. Also, many of these languages are facing obsolescence and extinction before documentation projects can take place.

There are many language documentation projects underway around the world, as well as many archives set up to receive, save and disseminate collected data (for examples, see the archives that participate in DELAMAN [3], and efforts like The Rosetta Project and PanLex Project [4] at The Long Now Foundation). Another approach is to develop a comprehensive digitization of every human language, enabling them for computational development and cross-linguistic comparison [5].

This is quite the herculean task and one that is quite possibly never complete. However the more progress we make, the better equipped we will be to understand alien communication, should we ever encounter it.

References

[1] The Long Now Foundation see http://www.longnow.org.

[2] The Rosetta Project and Rosetta Disk see http://www.rosettaproject.org.

[3] DELAMAN: Digital Endangered Languages and Musics Archive Network see http://www.delaman.org.

[4] The PanLex Project see http://www.panlex.org.

[5] See Steven Bird and Steven Abney. 2010. The Human Language Project: Building a Universal Corpus of the World's Languages in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.