

Building a Corpus of Comprehensible Text

by Greg Thomson

Used by permission of the author.

By the time native speakers have reached adulthood, they will have been exposed to thousands of hours of adult speech. They will also be a source of normal adult speech and will not usually worry about whether their own speech is similar to that of adult speakers in general.

Not so for the adult second language acquirer. It is quite possible for someone to develop a lot of speaking ability with a small amount of actual exposure to a language, and then do a lot of wondering. Analytically oriented language learners in particular can learn a stock set of structural frames and learn to substitute vocabulary into the various slots in those frames. With a vocabulary of a few hundred words, and a collection of a few dozen sentence patterns, and lots of experience at constructing sentences, a linguist/language learner can develop an exquisite capacity for expressing any imaginable meaning with amazingly limited resources. I know. I have been there.

A preferable scenario would have the adult language learner receive enough exposure to a language that s/he develops a more native-like feel for what is normal and what is abnormal. This may not be all that easy, unfortunately. The amount of exposure, like the amount of language acquisition, results from the complex interaction of two variables: the language learner and the social context. If the personality of the language learner is such that s/he prefers interlinearizing text to being with people, or if the social context is such that massive immersion in a speech community is impossible, then it will be desirable for the would-be language acquirer to come up with a strategy that will provide a modest amount of exposure to the target language. One strategy includes daily language sessions with a target language resource person (LRP) leading to the development of a corpus of fifty hours of tape recorded text with which the acquirer is well acquainted.

Early Texts (=Stage One)

Obviously, you are not going to get a feel for a language from listening to unintelligible texts. At the outset, you can hardly comprehend anything at all. By the end of the heavy language learning period you want to be able to comprehend just about anything you hear. The comprehensible corpus provides a partial means for moving from the nothing-at-all stage to the almost-everything stage.

It is well known that before one knows a language at all, utterances in it are only comprehensible insofar as the meaning of linguistic expressions is inferable from nonlinguistic information. Such inferences are possible, for example, if you can observe

what the speaker is talking about simultaneously, with hearing it talked about, or if the speaker is generous in the use of iconic gestures.

Of course, you can learn the meanings of expressions by means of translations from the mother tongue. However, there is a difference between knowing that *imitaa* is a translation for dog, and being able to recognize *imitaa* and relate it to a dog when some speaker uses it in a meaningful context. Finding out a translation equivalent can indeed be a step toward being able to attach significance to expressions used in real contexts. Nevertheless, I have nothing further to say about that approach to making input comprehensible. Instead, I am interested in making the target language more directly comprehensible in its own right, as a referential and interactional system.

So at the outset, you need visual aids if the input is to be comprehensible. During the first month or two of language learning it is possible to learn a lot of language through methods involving physical responses to instructions, warnings, predictions, and so forth. It is possible to learn a wide range of both vocabulary and grammatical constructions through such methods. For instance, person and number categories can be learned: “Do I have a pencil? Do you (plural) have pencils?” Specific constructions, such as conditional sentences can be learned: “If I have a pencil, then give me another one”; or temporal clauses: “When I look at her, hand me a cucumber.”

Such activities are typically conducted during dedicated language sessions. The early learner might spend an hour or two each day in such sessions with an LRP. At least an hour will be spent in planning and preparing for these language sessions. Another hour may be used in organizing the accumulating language samples for analytical purposes, and perhaps doing some analysis. During the sessions with the LRP, you will want to tape record the learning activities. You can then spend an additional two or three hours listening to the tapes, either responding physically once again, as in the actual session, or visualizing what was happening as the recording was being made. The act of recalling and visualizing in response to the audio recording is an important step. The language used in the session was comprehensible because it was scaffolded by the objects and actions in the immediate physical context. An essential feature of language is what is called displacement. This refers to the ability to use language to talk about situations which are not present, which are displaced in time or space. By listening to tapes of the sessions afterwards, you are relating the language expressions to memories, which is somewhat different from relating them to what is currently being perceived visually.

Along with physical response activities, you will also make use of photos and/or line drawings. These can be from books and magazines, but it is more effective to prepare one’s own photographs for this purpose. In this way you are able to illustrate a wide range of daily scenes and activities, and even to break down scenes into components, and activities into steps. Angela Thomson and I were able to snap over a hundred useful photos in roughly two hours.

Initially, the LRP will simply identify people in the pictures as men, women, boys, girls, people, children, young men, young women, old men, old women, and perhaps other

general categories of people based on categories like age and gender. The LRP can make such statements about scores of pictures right off the bat. "This is a man. These are some girls. Etc. Etc." This helps the LRP to warm up to the idea of using the pictures to scaffold communication. Furthermore, it provides relatively demanding, yet potentially comprehensible, text material for the first day or so of language learning.

Once it is easy to comprehend all such identificational statements about the people in the pictures, you can have the LRP make the most natural descriptive statement about what the people are doing in each picture. "This man is mending a net. This woman is cleaning her rifle. Etc. Etc." Now there will be a fair amount said which is not comprehensible to you. The next task is to make it comprehensible. To do this, you begin by having the LRP explain (in the contact language, assuming there is one) what s/he said in connection with each picture. This requires that you play the tape for the LRP. You make notes regarding the objects, actions, properties, etc. to which the LRP has made reference in connection with each picture.

Next, you focus on those objects, activities, or properties. The point is to learn to recognize the expressions in speech. The LRP can drill you by asking "Where is the net?" "Where is the rifle?" Etc. Etc. "Is there a rifle in this picture?" "Is there a net on this page?" Etc. Etc. In any such comprehension activity, learning proceeds most efficiently if initially there are only two choices, and then slowly, one at a time, new choices are added. Thus initially you only have to choose between a net and a rifle. As soon as this becomes easy a third item, say, a pail, is added. When it is easy to recognize all three expressions, a fourth is added, and so on, until about ten items have been added. For actions, the LRP can ask you things like "Who is washing something? Show me a picture of a woman who is cleaning something."

You may come up with other methods for becoming familiar with the vocabulary which has been used in the previously recorded picture descriptions. Now you can listen to the tape. Initially, you will want to listen several times with the pictures arranged in the same order as the descriptions on tape. Later they can be shuffled, and the challenge is to find the picture that is being described. Finally, keeping in mind the importance of displacement, you can listen to the tape without the help of the pictures, simply recalling the appropriate pictures mentally.

As time goes on, the pictures can be used to highlight particular constructions. For instance, the LRP can use a subject relative clause such as "Show me a woman who is cleaning a rifle. Where is the man who is mending a net?" in connection with each picture. Very quickly you will have the experience of comprehending a hundred subject relative clauses. The LRP can use object relative clauses in a similar way, either on a separate pass through the pictures, or on the same pass as the subject relative clauses. An object relative clause has the form "Show me the net which the man is washing. Where is the rifle which the woman is cleaning?" Thus in a relatively natural and communicative way specific grammatical constructions are repeated scores (or hundreds) of times.

In using the tapes privately for listening and recalling, you will listen to the tape made in the current day's session, and also to some made in earlier sessions, so that all sessions going back to the beginning are occasionally reviewed during the early weeks of full-time language learning. The sessions include both physical response activities and picture descriptions. Each week you can glean samples from every session which illustrate all of the vocabulary and grammatical constructions covered. The resulting gleanings should comprise about fifteen minutes from each week, or about an hour by the end of the first month.

Upping the Ante (=Stage Two)

A month has elapsed. You have only compiled an hour of material in the comprehensible corpus. The goal is fifty hours. Fifty months? Uh-uh. When you did not know any language at all, there was a need for a lot of repetition, not to mention weird texts which were tied to the here-and-now physical setting and participants, or to here-and-now pictures of the then-and-there.

Times have changed. You now recognize around a thousand lexical items, and can recognize the high frequency grammatical forms and constructions. The function of some grammatical morphemes may be unclear, but at least the forms themselves come through clearly. There may be important forms that have not been encountered yet, due to the restricted genres, but you have loads of language to work with.

The initial challenge was being able to recognize words and morphemes at all. That was the point of direct visual scaffolding of the verbal input. Full-fledged comprehension ability does not require that. So let's try some normal language. Have your LRP attempt to stay within the range of vocabulary and structure familiar to you, and tell you a story that is commonly told to small children in the target language community, but unknown to you. You will be able to understand it, right? Wrong (probably). So now what do you do? Understanding texts which are heavily scaffolded by visual input has become easy for you. Understanding a simple children's story is too difficult for you. Is there anything intermediate in difficulty between descriptions of pictures and simple children's stories?

I thought you would never ask. Indeed there is. Language comprehension involves working from evidence to conclusions. For the native speaking listener, it is often possible to conclude another speaker's intended words from phonetic evidence alone. You might even think that the native speaker simply hears the words, and need not conclude anything. That is not so. All that is heard is sound, and a lot of concluding must be done to get it into words and phrases. That concluding process is called parsing. The native speaker could overhear a snatch of that children's story, say a three-word phrase, and parse it into words on the basis of phonetic evidence plus a bit of grammatical detail. Not so with you. You may overhear a snatch of language containing only words you know and structures you know, and yet it is just a blur to you, and you have no idea that it contains those familiar words unless, perhaps, you capture it on tape and replay it several times, and, perhaps, get your LRP to say it more slowly. Your parser does not work very well yet.

What you were doing with visual scaffolding was helping your parser out by making meanings more predictable. You want to move beyond your need for visual scaffolding, but a simple children's story is too unpredictable for your parser to handle. What you want now is texts which, though not dependent on visual scaffolding, are nevertheless quite predictable. That is the intermediate level of difficulty we were looking for.

For example, suppose instead of a children's story from the target language community, your LRP were to tell you the story of Little Red Riding Hood in the target language. Now that is predictable. You may be amazed how easily you can understand it. Of course, you would need to first tell the story to the LRP in the contact language. At this stage of learning I once had an LRP tell me the entire Biblical story of the patriarch Joseph. Good story; I always enjoy hearing it. It was particularly enjoyable to hear it in a new language; it was surprising how much I could understand. And of course, there were expressions I could not understand.

The nature of the daily language sessions with the LRP changes at this point. For one thing, it is much easier to have longer sessions, say, two or three hours per day, with less preparation. There can be various sorts of text, as long as the content is fairly predictable for you. We have mentioned the possibility of being told a story which you already know. You may also have the LRP tell you of recent familiar incidents in the community. You may engage in a shared activity with the LRP, say a shopping trip, and then have the LRP tell a third party everything you did together. This should elicit a reasonably accurate text that is quite predictable to you.

This is also the stage at which the Series Method is most helpful. The LRP may tell you all the steps involved in a common activity, such as going to get water. "Going to get water" would be one series. "Setting a trap" might be another. "Getting dressed" might be another. It is possible to come up with scores of such series covering a wide variety of the activities and experiences of every day life. Such series can be told in excruciating detail. By their very nature, series texts are predictable, taking advantage of the predictable, episodic nature of recurring human experiences.

You now have a variety of methods for eliciting texts with the desired property of predictability. You are ready to go on building your comprehensible corpus. In your language sessions you will have the LRP tell you such predictable texts. As s/he tells them to you, you capture them on tape. There will be lots that you do not understand on first hearing. On the other hand, you will often be able to guess at the meanings of new words based on the context. Once the text has been uttered, rewind the tape recorder, and begin to listen to it together with the LRP. Every time you fail to understand something; ask the LRP to explain it to you as simply as possible (in the target language!). Make notes regarding the explanation of the parts you were unable to understand.

Once again, you will listen to the tapes when the LRP is not around. You will be surprised that you are often able to understand the newer vocabulary and constructions without recourse to the notes you made. As often as necessary, however, you will refer to

the notes. After you have listened to a text several times, it should no longer be necessary to refer to the notes. It is now part of your comprehensible corpus.

There continues to be a lot that you hear in every day life that you cannot parse. One reason for this (among others) is that your parser still has trouble just keeping up. You do better with slow, careful speech. Your parser will get faster as you use it more. Whenever you hear something that is 100% fuzz to you, your parser does not get exercised. The key to exercising your parser is to keep giving it lots of comprehensible input. You get comprehensible input from your comprehensible corpus, but you also need input that is fairly novel if you are to develop the ability to quickly parse novel material. Your language sessions are the most stress free opportunity for this, since you have a paid conversational partner who is precisely tuned to your current level of speaking ability. The process of conversation at this point is often described as the negotiation of meaning. Your ability to comprehend depends on the cooperative efforts of your LRP and you working together. Often s/he has to work hard to get a point across to you (and even harder to understand a point you are trying to make!).

You also have opportunity to negotiate meaning with conversational partners in ordinary social situations, such as visiting neighbors. This should be a part of your daily experience. It is good to bear in mind that communicating with you is hard work for target language speakers at this point. You will need to be sensitive to this. Better to make your visits too short than too long. That way people can look forward to your next visit rather than dreading it. If, on the other hand, you are not one to get out and visit, this will be an area for specific goal setting and self-discipline. Of course, if you are learning from one or more RPs in a situation where you do not have access to a speech community, you may have to postpone visiting until you can visit a speech community for some intensive socializing. Be encouraged that by methods such as those described here you really can develop basic conversational proficiency prior to visiting a speech community.

It is hard to specify a time limit for this stage of language learning in terms of some number of weeks or months. It will probably not be much less than two months, and may be three or four. Let's be optimistic, and say two. During this time, you should aim to add a half hour per week to your comprehensible corpus, being careful to review earlier materials at later points. After these two months (or whatever), your comprehensible corpus should be about five hours long. It will be rich in vocabulary and structure, though poor in naturalness and style.

Getting Serious (=Stage Three)

I mentioned that one reason you are often unable to parse what you hear at this point is that your parser is slow and inefficient. It is slow and inefficient in dealing with phonetic detail and lexical and grammatical information, and you often need to resort to guessing at what you are missing. Unfortunately, your ability to guess is also hampered. For one thing, you are struggling so intensely to process phonetic, lexical and grammatical information that you do not have a lot of mental resources left over for effective guessing.

It may seem unfair, but the easier parsing becomes, that is the more you are able to efficiently use phonetic, lexical and grammatical information, the easier it becomes to guess! Such an increase in parsing efficiency will come with time.

There are other problems that make guessing meanings difficult for you at this stage. Even native speakers must make use of inferential processes that go beyond the information given, to the information intended. If my son sees me drinking a can of soda and says "I'm thirsty", I know he will not be too pleased if I suggest a glass of water. Any narrative describes a sequence of events that contained infinite detail in the real world. Only a small portion of the real-world details are mentioned in the narrative. The listener fills in a lot of detail, while remaining indifferent to other details.

For the second language speaker (i.e. you), drawing essential inferences is not always easy. For one thing, successful inference drawing depends critically on the speaker and listener sharing a common background of knowledge. Much of this knowledge is specific to the local culture. This is sometimes referred to as the cultural knowledge bank. Successful communication depends on the shared cultural knowledge bank. Unfortunately, you do not yet share it.

Another reason that drawing essential inferences is difficult at this stage is that the second language learner misses all sorts of clues as to where s/he is in a discourse. These clues are found in things like discourse markers (e.g., "low and behold" warns the hearer that something unexpected is about to occur) and in the overall structure of the discourse. Have you ever told someone "You've lost me"? You were hearing the words. but you felt you were missing the point. In a second language, missing the point often mean missing the words as well. And it happens quite often.

It really comes back to the same problem we faced from the beginning. You are still unable to take advantage of the predictability of speech. Initially you could only overcome that problem through having visual means of making speech predictable. Then during the previous stage (stage two), you depended on your knowledge of what was being talked about to make the speech predictable.

Now you are running up against a barrier. To native speakers, what makes speech predictable, hence comprehensible, includes the shared knowledge bank and discourse clues. The methods employed thus far have help you to continue developing your general parsing ability. You now have a substantial linguistic basis for comprehension. However, you are far from sharing the cultural knowledge bank. and from being able to take advantage of discourse clues. You need ways to overcome both of those limitations.

James Spradley (1979) virtually equates doing ethnography with doing language learning, and at the stage of the game I am now focusing on (stage three) I agree with him. Further progress in comprehension ability depends to a significant degree on acquiring cultural knowledge. Certainly, much has already been learned, but it is only a start. It is time to take the cultural bull by the horns.

Language sessions with the LRP now center around ethnographic interviewing. This is not the place to describe Spradley's method in detail. In brief, you will have been compiling a list of social situations. A social situation is a recurring state of affairs that can be defined in terms regular participants, locations, props, etc. Two neighbors meeting at the village well might be a social situation. An interchange between a villager and a traveling merchant with a wagon of wares might be another. There are hundreds of social situations which can be identified in any culture. You will use your list of social situations as the basis for conversations with your LRP. You may ask a question about a typical instance of a social situation ("What goes on when a merchant comes to a house in the village?"), or about a specific instance ("Tell me all about the last time a merchant came to your house"). Such questions covering an entire situation are called grand tour questions. The responses are tape recorded and added to your comprehensible corpus. You will go over the tapes, identifying incomprehensible details, and clarifying them, making relevant notes. You will listen to the tapes privately until they are easy to comprehend.

It may seem as though this is just more of the Series Method. It is fundamentally different. The Series Method was an artificial method. If someone told you all the steps in winding a watch s/he was telling you something you already knew. That is unlike normal conversation. It is a kind of conversation that only occurs when someone's purpose is to help someone else learn how to say things. Normal conversation is not aimed at helping one of the interlocutors learn how to say things! Rather, it is concerned with conveying facts and establishing or strengthening social bonds. Ethnographic interviewing is thus a normal form of communication. You are repeatedly reminding the LRP of how much knowledge you lack about the locculture, and using the interview, as a means of gaining knowledge that you really want. You are getting serious about being an ordinary speaker of the language. The Series Method was also artificial in terms of the excessive amount of detail that went into the description of activities. In ethnographic interviewing the amount of detail given is appropriate to the communicative purposes of the interchange.

Often the response to a grand tour question reveals subcomponents which form the basis for mini-tour questions. The grand tour question may deal with the whole wedding. A mini-tour question arising from it may deal with the ritual washing of the groom, which is one part of the wedding. The responses to mini-tour questions are taped and added to the comprehensible corpus.

The grandest grand tour question is "Tell me what happens through the course of a typical lifetime". A specific form of the question would be "Tell me the whole story of your life as you can remember it." This can generate a lengthy text indeed. This is an important kind of text. The shared knowledge bank of the community is based on lifetimes spent together. The only way you can get at that kind of knowledge is through interviewing. Of course, you will want a variety of RPs of various ages and backgrounds for this purpose.

Within the life histories elicited will be ample opportunities for mini-tour questions, and responses to mini-tour questions can prompt sub-mini-tour questions. All responses are

taped, and the incomprehensible parts are clarified, and you listen to the tapes privately until comprehension is easy.

There is a lot more to Spradley's method than I have described here. Much of the time spent with your LRP, perhaps a few hours per day now, can be devoted to these other activities. In the present context, the main concern is that you learn to comprehend texts with ease.

Another important basis for conversations with an LRP has to do with learning social skills appropriate to the target language community. You have now been interacting with the target language group for many weeks, at least. You have experienced points of friction between you and members of your host culture. How do I know? I just know, that's how. Furnham & Bochner (1986) have suggested noting any recurring sources of cross-cultural strain. You can explain the type of situation to your LRP. Tell the LRP what happens. Describe how you tend to respond. Ask how s/he would respond. Tape the LRP's response to such questions, and add them to your comprehensible corpus. Again, the communication is authentic in that it is fulfilling a serious need for the transfer of knowledge. It is a special kind of discourse, heavily influenced by the fact that the LRP is speaking to an outsider with limited speaking ability. Still, that is a real speech situation (even if it has never existed in the community previously).

During this period you will do well to train one or more native speakers to write, if the community is not already literate, and hire them to transcribe your ethnographic texts for you. Having the texts in a written form and reading them in addition to hearing them can provide a powerful reinforcement to your aural language exposure. It is not a good idea to transcribe them yourself, since you will quickly find yourself doing nothing else. At this stage of language proficiency, transcribing texts is arduous. It gets easier as your proficiency in the language increases, but it is easiest for a native speaker. Training one or more native speakers to write does not take you away from language learning, since it provides opportunity for extensive conversational interaction. Transcribing texts does, to some extent at least, take you away from serious language learning. There will be some benefits, to be sure, but the growth of your comprehensible corpus will slow to a snail's pace, and you will switch your focus away from developing your ability to deal with spoken language as you hear it.

So far I have talked about ways you can increase your comprehension ability by coming to share more and more of the cultural knowledge bank of the community. There is also the need to be able to use discourse context to make language more comprehensible. In general, only a very advanced language learner is able to make use of discourse context in a way similar to how a native speaker does it. Presumably many of the discourse cues and structural features involved are acquired late. Nevertheless, discourse can be an important aid in making input comprehensible. I recall an advanced language learner describing how he had difficulty understanding what was said in an interchange if the native speaker only spoke three or four sentences. He said that in a longer conversation he was able to key in on the topic and start comprehending. I could easily relate his experience to some of my own. I noted that at the stage of language learning I am now

discussing, if I came into church in the middle of a sermon, I could not effectively follow what was being said. However, if I was present from the beginning of a similar sermon, I would be able to effectively follow the content right through the middle and to the end of the sermon. There is something to flowing with a train of thought. Similarly, I found it difficult to follow the plot of a television drama. However, if an LRP carefully explained to me what was happening in the early part of the plot, I found it much easier to follow the rest of the drama without the benefit of such explanations.

The phenomenon I am describing opens an important new possibility for the development and use of the comprehensible corpus. Suppose a text is recorded, let's say a folk tale, which seems largely incomprehensible. The approach to the comprehensible corpus which I have described so far would have you go over the entire text with the LRP, clarifying the difficult portions. In many cases now, a better approach will be to have the LRP explain the early part of the text in simple language that you can understand, but leave the latter portion unclarified. Your goal is then to comprehend at least the main gist of the entire text through repeatedly listening to it on your own. You can subsequently check with the LRP as to how accurately you have understood it.

The stage of language learning I have been describing in this section could go on for several months. You should aim to add about fifteen minutes to your comprehensible corpus about four times per week, or in other words, four or five hours per month. Recall that in your first month you only preserved an hour of comprehensible text in your taped corpus, and in the second month (or so) two hours. After six more months, then, you will have close to thirty hours of comprehensible text on tape. For some people, this stage should continue for considerably longer than six months.

I have said almost nothing about the production side of language learning, which is crucial to a balanced view of the whole experience.

The section that follows assumes that the reader is a serious field linguist. At this point, some readers may part company with us. Before saying goodbye let me at least say that before you discontinue full-time language learning, you will want to make sure of certain things. You should be able to achieve any communicative need you have in the target language without undue agony. You should have a rich social life within the target language society. You should have a work life which exposes you to the language constantly and requires you to respond in it. In any event, you will probably want to continue with a full-time focus on language and culture acquisition for a total of eighteen months to two years. If the language has a written form, massive reading will be an important component. If there are films and television programs in the language, these can be a tremendous aid. If you are dealing with a preliterate language, these options will not be available. The only option for the more advanced stages of language learning will be full-time shared lives.

Scaling Everest (=Stage Four)

A field linguist typically collects a corpus of language specimens. That is where I got the term “corpus” in the phrase “comprehensible corpus”. So far our comprehensible corpus is not always of the quality that might go into a linguistic corpus. The texts from the first weeks which had your LRP telling you things that required your physical response may provide isolated examples of linguistic phenomena, but they involve a relatively inauthentic use of language. The descriptions of pictures may provide somewhat more authentic language, once you reach the stage where the LRP has a lot of freedom in saying whatever s/he feels are the most sensible things to say in connection with each picture. Your Series Method texts are certainly low in authenticity and naturalness. Your ethnographic texts are on a different level. They are valuable for the information they contain. The language used in them may be of a special form, since the speaker had to modify it in such a way that you would have a chance of comprehending it. Furthermore, the speech situation of an ethnographic interview is not something which occurs normally within the target language community. So although the ethnographic texts should certainly be transcribed (by a native speaker, of course), translated, and archived, they may not be your best texts for linguistic analysis. You have thirty hours in your comprehensible corpus. You can randomly choose any tape without looking at the label, start playing it at a random spot somewhere in the middle, and understand what you are hearing. That is why it is called a *comprehensible* corpus. Thirty hours of language, rich in vocabulary and cultural domains, and representing every basic grammatical construction—and you have enough language knowledge in your own personal brain to be able to understand any of it. And you have only been at it for eight or ten months. Not bad (even if you have actually been at it for fifteen or eighteen months).

But you have yet to add another twenty hours to your comprehensible corpus. Perhaps a fair bit of that will continue to consist of ethnographic texts. But at least ten or fifteen hours of it should consist of language specimens which are as authentic as possible, and which represent a variety of speech situations and registers.

Recording authentic language samples is not easy. Clandestine recording would be the most straightforward method, but (UN)fortunately it is not ethical. As soon as the speakers are aware that their speech is being observed, there is a loss of authenticity. Introducing a microphone may lower it even further. If the speech serves no communicative function other than to provide a specimen, it cannot be truly authentic.

But then, why be a perfectionist? Years ago some field linguists had native speakers dictate lengthy folk tales, a sentence at a time, while the linguist transcribed everything. The resulting written texts were not all that authentic, but were extremely valuable just the same.

To be fully authentic, a text needs to be spoken solely because the speaker feels a need to communicate something to the hearer, without any thought of it being preserved. A folk tale told at the fireside for the entertainment of the listeners is authentic. A speech made to an audience for the kind of purposes for which speeches are normally made is authentic. Those are examples of authentic monologues. All of the sorts of interactions that take place in the normal course of life in the community provide instances of

authentic dialogues. The problem is, it is difficult to make good quality recordings of such authentic speech. So we tend to fall back on recording texts in formal sessions which we set up for that purpose. Certain principles will help to raise the authenticity and quality of these texts.

If a monologue is to be told, the audience should be native speakers. The communication should be motivated, if possible. For example, a folk tale may have the motivation of providing entertainment, or reinforcing certain beliefs or values. It will probably not involve the speaker conveying new information to the audience. Most other types of monologue do involve the transmission of new information. The speaker has something which s/he wants the hearer to know, which hearer does not know, but wishes to know (or at least feigns to wish to know). That is what is meant by saying that a text is motivated.

Linguist Austin Hale has pointed out (p.c.) that to learn what distinguishes effective communication from less effective communication, it is necessary to examine text produced by effective communicators. He points out that different speakers may be effective for different registers or speech situations. He points to a variety of factors that are essential to comprehending a text. Certainly the better you know the speaker and his or her views, and are familiar with the subject matter, and the knowledge bank of the hearers, especially as it relates to the topic at hand, and know the reason for the encounter in which the text occurs, and the purpose of the communication from the perspectives of the speaker and the hearers, the better chance you have at accurate understanding, (That sounds like a tall order, but in fact, it is the sort of approach to text understanding which is required of a good translator—see Schweda-Nicholson. 1987).

Thus Hale recommends that along with the text, a record be preserved which attempts “to characterize as accurately as possible what the speaker was doing, or hoping to accomplish through the discourse.” It is also important to note the audience and the quality of their interaction with the speaker. In the early stages of the development of your comprehensible corpus, you were the audience. That was important if the speech was to be comprehensible to you at that stage. But you want to push on to the stage where natural speech addressed to native speakers is intelligible to you. To do that you want a large quantity of text which was addressed to native speakers.

Hale suggests that a good quality of discourse is likely to result if the speaker feels “challenged to prove something, or to share something about which he or she [has] deep convictions”. Such a situation can exist even in the case of narrative texts. For example, a story told by an elder to young people might have the purpose of proving that life was tough in the olden days.

It may be that the level of quality I am describing is difficult to achieve, especially if, in addition to quality texts, we desire quality voice recordings (or even video recordings). Hale points out that it can be an advantage to train a member of the target language community as a specialist in text collection. Short of that, he suggests you at least attempt to get the speaker to speak *as though* specific authentic conditions were in force.

There is much more to be said about the development of a corpus of text for linguistic and analytical purposes, but our topic here is the development of a comprehensible corpus as a language learning method. In previous sections I spoke in terms of getting fifteen minutes of text recorded four times per week, and using a lot of language session time to go over the tapes with the LRP to make certain that everything is comprehensible, and correctly comprehended. At the stage I am now describing, the methods of text collection and varieties of recording situations are less structured and less predictable. The texts will commonly be spoken by a person different from the LRP who assists you in understanding the difficult bits. Some weeks you may have opportunity to obtain far more than an hour of text. The important thing from the standpoint of your comprehensible corpus is the rate at which you work through the text. For this you can take a structured approach of adding fifteen minutes to the comprehensible corpus four times a week. It will thus take four or five months to add the final twenty hours to your comprehensible corpus. Of the twenty hours, you may wish for ten to be high quality monologue, and ten to be more average quality conversational discourse.

It is also a concern during this advanced stage to develop an understanding of the written regi. The odds of your achieving native-like speaking ability are not very high. A more realistic goal for you in relation to spoken language will be to achieve excellent communicative competence with some imperfections of accuracy and usage when speaking under pressure in real time. However, you have a much better chance to learn to compose excellent written discourse, since writing is done without the demands of real-time processing, and you can apply your explicit knowledge of everything from morphology to discourse during the production of written compositions. If the target language has a long written tradition, a written register will have developed. If the language is just beginning to be written, then the written register is yet to develop, and you will be an observer of the process. In that case, during the advanced stage of language learning you may be able to help foster the development of vernacular literature. One of your important contributions to the target language community can be to open this possibility. In this way you can develop a corpus of written language which was intended as written language. This can be read aloud as well, and tape recorded as a source of comprehensible input in the aural modality. In any case, you will want exposure to a large amount of high quality text, as well as more typical colloquial speech, and you will want to include ten or twenty hours of such text in your comprehensible corpus.

Once again, native speakers should be employed to transcribe oral texts that have been tape recorded. As with your ethnographic texts, you will want to provide free translations to go with the transcriptions of the ten or twenty hours of high quality text you have captured. A smaller quantity of these transcriptions, perhaps twenty-five or thirty percent, will be given an interlinear translation, and analyzed grammatically in a variety of ways.

Was it worth it?

I will be bold and state categorically that it was worth it. You have spent a year and a half devoting quite a bit of your time to constructing and reviewing this comprehensible corpus. If you are living in an immersion situation, where you are required to interact in

the language in all areas of life all of the time, then you will have been exposed to many hundreds of additional hours of speech. But your comprehension corpus gave you a way to systematically and steadily increase your ability to comprehend speech. It increased the percentage of those hundreds of hours of natural speech which were comprehensible to you.

On the other hand, suppose your opportunity to be involved in real speech situations was restricted. Maybe you only had access to a couple of speakers. Then the development of your comprehensible corpus has been central to your acquisition of the language. How else would you have developed comprehension ability to the extent that you have? (You have developed speaking ability as well, but that is another topic.)

Some language learning situations are more challenging than others. In the least challenging situations some people will learn a language quickly without even concentrating on language learning. But in other situations, and with other people, successful language learning may take a lot of concentrated effort. Occasionally I hear of someone who “just can’t learn languages”. I say give me that person, a couple of native speaking RPs, and eighteen months, and I will show you a successful language learner. The comprehensible corpus cannot be the whole answer. But it could be a whole lot of the answer.

References

- Furnham, A., & Bochner, S. (1986). *Culture Shock: Psychological Reactions to Unfamiliar environments*. London and New York: Methuen.
- Schweda-Nicholson, N. (1987), Linguistic and extralinguistic aspects of simultaneous interpretation. *Applied Linguistics*. 8(2), 194-205.
- Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart and Winston.