Making a Tibetan Speech Corpus

A Beginner’s Guide to Collecting & Transcribing Speech Data
Introducing Corpus Linguistics

0.1 What is Linguistics?
  0.1.1 The Scientific Method 4
  0.1.2 Linguistics 6
0.2 What is a Corpus?
  0.2.1 Building a Model 11
  0.2.2 Using Machines 13
0.3 Uses for Corpora
  0.3.1 Frequency Lists 15
  0.3.2 Readability & Plain Language 16

1 Project (项目) -- Creating a Speech Corpus 20
  1.1 The Team (Overview) 20
  1.2 Project-Level SOPs 20
    1.2.1 General Manager SOP 20

2 Person (人员) -- Collecting People's Speech Data 23
  2.1 Data Collector SOP 23
    2.1.1 Recording Equipment 23
    2.1.2 Recording Metadata 24
    2.1.3 Recording People's Speech 25

3 The Sessions (任务) -- Transcribing Speech Data 31
  3.1 Transcriber SOP 31
    3.1.1 Pre-Processing the Speech Data 31
    3.1.2 Transcribing 39

4 What Next? 48
  4.1 History of Corpora
    4.1.1 What does history tell us? 50
  4.2 Analysis Tools
    4.2.1 AntConc 51
    4.2.2 Voyant 52
    4.2.3 SmartCAT 53
    4.2.4 SketchEngine 54
  4.3 Training Others 54

GUIDELINES for the PROJECT Metadata (项目) 55
  1 Types 55
  2 Formality 55
GUIDELINES for the PERSON Metadata (人员) 56
  1 Age 56
  2 Sex 56
  3 Birthplace 56
  4 Education 56
  5 Occupation 56

GUIDELINES for the SESSION Metadata (任务) 57
  1 Setting 57
  2 Audience 57
  3 Communicative Aim 57

Guidelines for Session Transcription (任务) 59
  1 Spelling 59
    1.1 New Words 59
    1.2 Numbers 59
    1.3 Abbreviations 59
  2 Disfluent Speech 59
    2.1 Grammar Mistakes 59
    2.2 Filler Words & Interjections 60
    2.3 Partial Words & Partial Sentences 60
  3 Other Markup 61
    3.1 Multiple Speakers 61
    3.2 Unclear or Unintelligible Speech 61
    3.3 Other Non-Words 61
    3.4 Anonymizing the Transcript 62

The Person Metadata Form - পুরোৱেৰ মন্তব্য বুখানাৰ মাধ্যমে ইংৰী গৈ ৬৩
Introducing Corpus Linguistics

0.1 What is Linguistics?

In this guide, we will teach you what a speech corpus is and how to make one. But first, you should know that corpus-creation is a linguistics tool. That’s why it’s helpful to know a bit about what linguistics is (and isn’t) before we start.

Briefly put, linguistics is the scientific study of language. You may remember “studying language” during your school days. You may have learnt things like grammar and spelling—the “right” way to use a language. Linguistics isn’t this kind of language study.

In linguistics, we aren’t worried about what’s “right” or “wrong”—instead, we are worried about studying a language as it is. This is what makes it scientific. Linguistics is making observations about language, and gathering data to come to conclusions.

0.1.1 The Scientific Method

What is science?

You might be familiar with logic. The classic example in Tibetan is: “ཕུན་ཚུགས་སོགས་ལེགས་པ་ནི་བོད་དབང་གི་དོན་དེ་ནི་།”—in the West, the classic example is: “Socrates is a man; all men are mortal; therefore, Socrates is mortal.” These are examples of deductive reasoning. Scientific thought, on the other hand, is built on inductive reason.

For example, how do we know that “all effects are impermanent” and “all men are mortal”? ... It’s because we’ve observed causes and effects and living beings over and over again and seen it. They never last. They always end. It is a premise based on experience, or observation (or, more accurately, very many observations).

All the scientific method does is make this process of “observing things” very clear, and very organized. That way, we can use the same method on more and more subtle things. By working very hard, and keeping very good notes of lots of data, we are able to come to conclusions about the nature of our reality.

In short, the method requires us to: 1) Observe; 2) Hypothesize; 3) Test.

1. Observe. Notice something about the natural world.
2. **Hypothesize.** Take a guess about why or how.
3. **Test.** Gather data. Analyze the data. Come to a conclusion.

*Let’s take an example to see how it works.*

**Fish Jump**

![Fish jumping out of the water]

**Step 1: Make an Observation**

Imagine you live near a lake. At some point, you’ll probably observe a fish jumping out of the water. We’ve just taken the first step in the scientific process: We’ve **made an observation**.

**Fish jump.**

**Step 2: Hypothesize**

You might begin to wonder, “Why do fish jump?” The next step in the scientific process is to make a guess that explains the phenomenon “fish jump.”

**Q:** Why do fish jump? Write down 2-3 guesses:

1. **Maybe fish jump because...**
2. **Maybe fish jump because...**
3. **Maybe fish jump because...**

**Step 3: Test**

To help solve the problem of why fish jump, we need to **collect data**. But what kind of data will help us answer the question? Actually, there are a lot of smaller questions that might help lead us to an answer. For example, we could test:
• **When do fish jump?** — sitting by our lake, we could note how many times fish jump, and at what time. (Do fish jump more at night or during the day? Do they jump more in sunny weather or in the rain?).

• **Where do fish jump?** — we could draw a map of our lake, and note where the fish jumped. Do they jump by the shore? Or in the middle of the lake? Both?

• **What kinds of fish jump?** — we might watch each jump for *what kind* of fish jump. Are they big fish, or little fish? What species of fish are they?

**Q:** Can you think of more data we could **collect** on fish jumping in our lake?

You might notice that in each example above, the more data we collect, the better. We probably cannot answer our question, “Why do fish jump?” by observing one fish jump one time, or even two jumping twice. But if we collect lots of data, about lots of fish, jumping lots of time, we’ll get closer and closer to coming to an understanding of why fish jump.

*Now, how does this relate to language?*

**0.1.2 Linguistics**

As we said before, **linguistics** is not just “studying grammar.” Instead, imagine a lake-full of language, swimming with words. Just as we can observe fish swimming and jumping, we can observe words being spoken (or written). The words people are using, the places they’re using them, and who they’re using them with, are all forms of **language data** we can collect.

Like our fish, we want to investigate not just one person’s writing or speech, but as much language as possible. We can investigate language by 1) **observing** it; 2) **hypothesizing** about it; and 3) **testing** our language knowledge by **collecting** and **analyzing** language data. If we have a question about language, we can collect data on how it’s used, when, where, and with whom.

**Language Change**

Let’s take a closer look at an example: the English second-person pronoun “you.” Some 500 years ago, English “you” was *exclusively* a plural pronoun (םְלֹא). When English-speakers talked about one person, they used “thou” (םְלֹא). In those days, many grammarians thought it was wrong and improper to say “you” to just one person.

Yet languages, like any phenomena, are impermanent. They change and grow, just like anything else. The word “you” didn’t stay the same. It first became the polite way to talk to
someone of a higher social standing; “thou” continued to be used for friends and family. Eventually, “you” completely replaced “thou.”

<table>
<thead>
<tr>
<th>15th Century:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominative</td>
</tr>
<tr>
<td><strong>Singular</strong></td>
<td>thou</td>
</tr>
<tr>
<td><strong>Plural (and formal singular)</strong></td>
<td>ye</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15th Century:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominative</td>
</tr>
<tr>
<td><strong>Singular</strong></td>
<td>you</td>
</tr>
<tr>
<td><strong>Plural</strong></td>
<td>you</td>
</tr>
</tbody>
</table>

Look at the table above: Is there any difference between singular “you” and plural “you” today? How do English speakers know if “you” means one person or many? The truth is, a grammar expert, not a linguist, probably made the table above. While no single version of the plural “you” is accepted as standard, or grammatically correct, native speakers do differentiate between “you” singular and “you” plural.

In America, for example, the main two plural forms for “you” are “you guys” and “you all / y’all.” Some linguists collected data from native speakers across America (by asking thousands of people, “How would you address a group of two or more people?”). This showed that people in the North say, “you guys” and people from the South say, “y’all,” and almost nobody at all only says “you:”
What makes this linguistics work is that:

1. An observation was made about language (English singular and plural “you” are the same!)
2. A hypothesis was formed to explain the observation (Maybe English speakers use a different word for the plural)
3. And data was collected to test the hypothesis (English speakers do use plural forms for “you!” They use “you guys” and “y’all.”)

If we are linguists, we care about describing language as it actually is, not as we think it should be. Let’s re-make our table from above, including the plural forms of “you” as they actually exist (as proven by our data):

<table>
<thead>
<tr>
<th></th>
<th>Nominative</th>
<th>Oblique</th>
<th>Genitive</th>
<th>Possessive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Singular</strong></td>
<td>you</td>
<td>you</td>
<td>your</td>
<td>yours</td>
</tr>
<tr>
<td><strong>Plural</strong></td>
<td>you guys</td>
<td>you guys</td>
<td>you guys’s</td>
<td>your guys’s</td>
</tr>
<tr>
<td></td>
<td>y’all</td>
<td>y’all</td>
<td>y’all’s</td>
<td>y’all’s</td>
</tr>
</tbody>
</table>

In English, “thou” (singular) became “you” (singular); “you” (plural) became “you guys/you all” (plural). What is a language change we can observe in Tibetan?

The important thing to note is that questions like, “Is it okay to use ‘you guys’ as the plural form of ‘you’?” or “Is it wrong to use ‘you’ for one person?” are not questions for linguistics. The reason is simple: we cannot collect any data that will answer that question.

Asking “right or wrong” questions like this about language is like asking, “Is it wrong for the fish to jump in the water?” All we can do is observe when it happens, come up with theories why it happens, and then collect data to test our theory.

**Applying Our Knowledge**

Finally, in linguistics, we’re also concerned with the practical applications of our research. How can we use our knowledge about a language in a practical way, to benefit people who learn or use language?

In our fish example, we can begin by collecting data. Let’s start with the question: **When do fish jump?** Again, we need to remember that if we just watch for one or two jumps, we can’t come to any conclusions. But, the more we watch, the more data we have. Let’s
pretend we sat at the lakeside for many days, and counted each fish jump, and recorded what time it happened at:

![Graph of number of jumps vs. time]

The more data we have, the surer we can be in our data. Here we can see, for example, that the fish jump a lot at two times of day: sunrise and sunset. Based on this conclusion, we can ask a new question: “Why do fish jump at sunrise and sunset?” Also think about how this might help us answer our initial question “Why do fish jump?”

**EXERCISES:**

1) Investigate why fish jump. Be sure to explain what data you’d gather to test them!

   1. **HYPOTHESIS:** Fish jump because...

      **TESTING:** To test this hypothesis, I’d gather data on...

   2. **HYPOTHESIS:**

      **TESTING:**

2) Now that you know when fish jump, tell me some **practical applications:** If you’re a fisherman, how can you apply your knowledge that fish jump at dawn and dusk?

   1. Since fish jump at dawn & dusk, if I were a fisherman, I would...

   2.
3) Can science answer a question like: “Is it morally acceptable to be a fisherman?”

☐ YES
☐ NO

4) How about a question like: “Why are there fewer fish this year than last year?”

☐ YES
☐ NO

5) What if we asked the question: “What is the most-used letter in Tibetan?” We will get a lot of opinions. Some people will say “ζ,” Other’s will say, “κ.” Still others, “η” or “ς.” We could debate all day without coming to any conclusion. How do we find the real answer?

Design a **scientific method** for finding out which letter in Tibetan is used the most:

What’s your guess? What’s the most frequently used letter in Tibetan?

**Key Concepts:**

- **Linguistics:** A *scientific* approach to language.
- **Scientific Method:** A process of: Making *observations* about a natural phenomenon; coming up with *hypotheses* to explain those observations; testing those hypotheses by *collecting data*. 

0.2 What is a Corpus?

Now that we know a little about the **scientific method** and what **linguistics** is in general, let's introduce the **corpus**. Simply put, a corpus is a large set of language data. (If we are applying the **scientific method** to language questions, it is **Step 3: Collect Data**). It can include any kind of language data:

- **Speech**: conversations, speeches, dialogs from movies or TV, song lyrics, etc
- **Writing**: poems, novels, religious texts, plays, etc

It might also include many kinds of **metadata** (data about data), like:

- **Who**: who spoke or wrote the words? How old are they? Are they male or female? Where are they from?
- **What**: what is the topic or genre? The dialect?
- **Where**: where was the speech spoken, or the text written?

Not all corpuses will include all of these things. What is included in a corpus will depend on: 1) what questions the corpus-creator is trying to answer or 2) what applications the corpus-creator has in mind.

**Q**: You want to create a corpus to see how language changes over time.

1) **What kinds of data** will you include?

2) **What kinds of metadata** will you include?

0.2.1 Building a Model

To recap, the **Scientific Method** requires us to:

1. **Observe**: Notice something about the natural world.
2. **Hypothesize**: Take a guess about why or how.
3. **Test**: Gather data. Analyze the data. Come to a conclusion.
As we said above, **building a corpus** is **Step 3** (and only Step 3) of the Scientific Method—**collecting data**. The kind of data we collect depends on what questions we want to answer (**Steps 1 & 2**). But what if we want to be able to ask a lot of different questions? What if we want a corpus for general use? A corpus that many people can analyze and test their different hypotheses on?

Let’s start by imagining the **ideal corpus**. An ideal corpus would include all the language data of all the speakers and writers of a language over the entire period of that language’s existence. If such a corpus were possible, it would mirror reality precisely, and our **observations** of the corpus would be equal to **observations** of reality. Anybody could test any **hypothesis** against the corpus, and get exactly the right answer!

Of course, it’s impossible to collect all the language of all the people. But, the closer we can get, the more our corpus will reflect reality. In short, this is the goal of a **representative corpus**—a collection of language that spans all the different kinds of language. A representative corpus will be sure to collect all the kinds of data (natural speech, prepared speech, writing) over all the kinds of metadata (different ages, sexes, locations, etc.) that are possible to collect.

In short, a corpus for **general use** is called a **representative corpus**. It is basically a **language model**—a small replica of the language at large. To precisely model a language, we need to be sure we cover all the different kinds of **people** who use language, all the different kinds of **language** that they use, and all the different **places** that they use it. Then, when we’re **collecting data** and making the corpus, we need to be sure we have **samples** from all these categories.

**Q:** You want to create a corpus for **general use**.

1) **How will you make it representative?** What kinds of **categories** will you include? Make a list. Be as detailed as possible! (You might think of something we haven’t yet!)

Start with lists for things like **a) Kinds of People, b) Kinds of Language, & c) Kinds of Places**
0.2.2 Using Machines

Another important point about corpus work is that it’s very important for our language data to be machine readable. Computers make many things possible that would be hard, or impossible, to do manually, by hand. For example, counting up all the letters in a million word corpus and organizing them by frequency would take years for a human to do by hand. A computer can do the task in a matter of seconds!

For an example, let’s come back to our question: “What is the most-used letter in Tibetan?” What was your guess for the answer? What method did you come up with to find the real answer? One way would be to keep a list by hand. We could pick up a book, and start counting! Of course, it would take us a very long time. On the other hand, if we have digital text, we can ask the computer!

<table>
<thead>
<tr>
<th>ཁ</th>
<th>ག</th>
<th>ང</th>
<th>ཅ</th>
<th>ཆ</th>
<th>ཉ</th>
<th>ཊ</th>
<th>ཋ</th>
<th>ཌ</th>
<th>ཏ</th>
<th>ད</th>
<th>ཕ</th>
<th>བ</th>
<th>ཞ</th>
<th>མ</th>
<th>ཙ</th>
<th>ཚ</th>
<th>ཛ</th>
<th>ཛ</th>
<th>ཙ</th>
<th>ཚ</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>ད</td>
<td>ད</td>
<td>ན</td>
<td>པ</td>
<td>དྷ</td>
<td>པ</td>
<td>ན</td>
<td>པ</td>
<td>ན</td>
<td>ད</td>
<td>ད</td>
<td>ན</td>
<td>པ</td>
<td>པ</td>
<td>ན</td>
<td>པ</td>
<td>ཐ</td>
<td>ཐ</td>
<td>ཐ</td>
<td>ཐ</td>
</tr>
<tr>
<td>9.6%</td>
<td>7.9%</td>
<td>6.9%</td>
<td>6.4%</td>
<td>6.2%</td>
<td>5.7%</td>
<td>5.5%</td>
<td>5.1%</td>
<td>5.0%</td>
<td>4.7%</td>
<td>4.6%</td>
<td>4.5%</td>
<td>4.3%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.2%</td>
<td>1.6%</td>
<td>1.0%</td>
<td>1.9%</td>
<td>0.9%</td>
<td></td>
</tr>
<tr>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td></td>
</tr>
<tr>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td></td>
</tr>
<tr>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td></td>
</tr>
<tr>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td>བ</td>
<td></td>
</tr>
</tbody>
</table>

In this chart, the computer counted all the letters of all the texts in all the volumes of the Kanjur. It then told us exactly how many letters there were (69,362,831) and exactly how many times each letter occurred. " psycopg2" was the most-frequent letter, occurring 6,683,352 times. In other words, the letter “ psycopg2" makes up 9.60% of the Kanjur. One laptop can perform this kind of task in a few seconds. How long do you think it would take one person? Two people? A team of 100 people?

**Q:** How many people would you need to count 70,000,000 letters in a few seconds?
Post-Processing

Here, we also have to mention that machines can’t read language like humans can read language. They don’t understand the meaning of language; they can only perform operations on the data that is there. For example, the computer won’t know if an instance of “ Zika” means “if” or “sick”—unless we give it a strategy to tell the difference.

The three most important strategies that help machines read text data are 1) Segmentation; 2) POS tags; and 3) Lemma Tags. We’ll talk a little bit more about these strategies in the final chapter (Chapter 4, “What Next?”). For now, let’s briefly introduce them, and explain why they are important.

“Segmentation” is a strategy that tells the computer where words and sentences begin and end. Words and sentences are important to humans because they represent units of meaning—of a single “thing” in the case of words, and of a single “thought” in the case of sentences.

Since a machine can only analyze text data (without understanding it), we need to mark word-level meaning—traditionally with a “space” character, as used in languages like English—and sentence level meaning. Segmentation is especially important for non-spaced languages, like Tibetan (and many other Asian languages). Without segmenting the text, the computer won’t know where one word ends and the other begins.

“POS” stands for “Part of Speech;” it is a type of annotation we add to raw text to help the machine understand the difference between words that have the same spelling, but different meanings (like our “ Zika” example from above). Each word gets marked with a “tag” that tells us if that word is a noun, a verb, a particle, or what.

If POS Tags are important for telling the computer which words with the same spelling have a different meaning, Lemma Tags are important for telling the computer which words have a different spelling but a similar meaning. For example, “ Zika” and “ Zika” are different tenses of the same verb, but have completely different spellings. The only way the computer will know these words are connected is if we tag them!

Post-processing text data with Segmentation, POS Tags, and Lemma Tags are important for many applications. They are key for making a spell-checker for Tibetan; without them, search engines (like Baidu) won’t work well with Tibetan; and they’re also important for just about any other application we may wish to use corpora in...
0.3 Uses for Corpora

In our “fish jump” example above, we asked, “How is the data we collect useful?” When we found out that fish jump at sunrise and sunset, we decided that those would probably be good times to go fishing. The data we collected—“fish jump at sunrise”—was useful to us as fishermen—we can fish at sunrise, and maybe catch more fish.

Q: How is collecting language data useful? Do you have any guesses?

What can we do with corpora? As it turns out, corpuses are very useful for just about anyone who works with language. That includes: Authors, Journalists, Teachers, Textbook Writers, Dictionary Makers, and Academics, to name a few. How?

0.3.1 Frequency Lists

As we saw above, one of easiest kinds of analysis we can do by machine is count things. Above, we counted letters. But in a corpus, we can also count words. One of the most basic, and useful, types of analysis we can ask the computer to do is give us a frequency list: A list of all the words in our corpus, ordered by how frequently they appear.

This is useful because it gives us a very clear idea of which words are easy—the frequent ones, the common ones we see and hear and use every day and almost everybody knows—and which words are hard—the infrequent ones, the ones that are rare and specialized and aren’t used every day and most people don’t know.

<table>
<thead>
<tr>
<th>Tibetan</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>ཆ་</td>
<td>一</td>
<td>the</td>
</tr>
<tr>
<td>ཆུ (ནུ་ནུ་ནུ་)</td>
<td>在</td>
<td>be</td>
</tr>
<tr>
<td>མ་</td>
<td>有</td>
<td>and</td>
</tr>
<tr>
<td>ཆུ་</td>
<td>個, 个</td>
<td>of</td>
</tr>
<tr>
<td>བ་</td>
<td>我</td>
<td>a</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>རིག་</th>
<th>བོ</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>དེ་</td>
<td>達, 这</td>
<td>to</td>
</tr>
<tr>
<td>བོ</td>
<td>了</td>
<td>have</td>
</tr>
<tr>
<td>ས་</td>
<td>他</td>
<td>it</td>
</tr>
<tr>
<td>རིན</td>
<td>也</td>
<td>I</td>
</tr>
</tbody>
</table>

0.3.2 Readability & Plain Language

Applying frequency lists to help write and edit texts is call **readability**. When authors know what words are **easy** and what words are **hard**, it helps them use language that they know their audience will **understand**. Think about the job of a teacher or a children’s book writer, for example. If they are using language that’s too hard, the child won’t understand the lesson.

Or think about a newspaper article. The point of a newspaper is to communicate information to the reader. If a journalist uses difficult words and long sentences, that information will only be understood by very educated readers—ones who know difficult words, and are able to read complicated sentences. But if a journalist writes plainly, simply, and clearly, the information can be understood by anybody.

The goal of an author, in other words, is to write in a **readable** style, using words that his audience understands—**plain language**. Making children’s books easy-to-read helps children learn how to read; making news media and popular literature easy-to-read builds good reading habits, and gives people access to information. An author who is easy to understand is widely read; an author who is difficult to understand is only read by a few people.

Take a look at the graph below. If a writer writes at a 16+ grade level (university level), using very difficult words and complex sentences, less than 2% of the population can read and understand it! By writing at a grade level of 10-15 (high school), fewer than 25% of people can read and understand. But, if the language is simple and clear (6-9 grade level, middle school), well over half of the population will be able to read and understand the text:
Q: According to the graph, what grade level do you need to write at to reach 90% of the population?

Professional Readers vs. A General Audience

We also need to understand why people read. People who don't read (and write) for their profession need a reason to read and write in their free time. If reading material is too difficult, it becomes like another job. If reading is hard and frustrating, people who don't need to read won't read. Or, they may end up reading and writing in another language instead, one that is easy to read and write.

In other words, most people are not academics, monastics, historians, or literary experts; they are not professional readers. They may never be. They will learn to read if and only if it's easy, convenient, and enjoyable. They'll read to get information from the news, or to relax and enjoy a story. Writers need to learn how to write for their audience; the tools corpuses provide are perfect for this kind of work.

Q: Do you read and write? In what language? Why? What do you use reading for? What do you use writing for?
Speech is the Foundation to Reading

It’s also important to know that there is a relationship between speech and writing. Written words symbolize sounds. When a child first learns to read, for example, they begin by making connections between the symbol on the page and the words they already know. And, whenever anybody reads, they do so by decoding sounds; it is these sounds that are connected to meaning.

By making a speech corpus, then, we are collecting the data that experts can use to address important language issues, like literacy. We are making a tool that can be used by authors, journalists, dictionary makers, and just about anybody else that works with language. And, our goal is to benefit the speakers, readers, and writers of the entire speech community.

Q1: How does a speech corpus help with literacy?

Q2: We want to create a corpus for the Tibetan language. How are we going to maximize its usefulness?

Key Concepts:

- **Metadata**: Data about data. For language data, we can ask: Where did it come from? Who spoke it / wrote it? When?

- **Representative Model**: Something that models, mirrors, recreates, or represents, a real phenomenon as closely as possible.

- **Practical Applications**: Using theories or models to provide a real, tangible benefit; making scientific research useful.
Project 项目
1 Project (项目) -- Creating a Speech Corpus

Now that we know a little bit about what a corpus is, and what it is used for, let’s get into the details of how to make one. The first thing we’ll need to do is gather the details about the project in general: Which language are we collecting data on? And what type of corpus are we making? Who is doing what work, and how will the corpus team be organized?

1.1 The Team (Overview)

Again, a corpus is a collection of machine-readable language data. So, we’ll need 1) data collectors, people who actually collect speech data (by recording it); and 2) transcribers, people who listen to the recordings, and type the conversations into the computer (so that it’s machine readable). The team will include:

1. Manager(s) - In charge of understanding and ensuring data quality & standard operating procedures (SOPs); will communicate these standards with the data collectors and transcribers, and then check their work using the Checklists
   a. Managers should have a deep understanding of the details of all jobs in Chapters 1-3, along with the goals of the project at large [Introductory Chapter], as well as the guidelines, checklists, and SOPs
2. Data Collectors - Will collect speech data (recordings) of representative samples of the Tibetan-speaking population according to the guidelines (more details on the job of Data Collectors is found in Chapter 2) and complete all the tasks on the Data Collector Checklist
3. Transcribers - Will transcribe speech data according to the guidelines (more details on the job of Transcribers is found in Chapter 3) and complete all the tasks on the Transcriber Checklist

1.2 Project-Level SOPs

1.2.1 General Manager SOP

The General Manager will be in charge of working with the team on a daily basis to make sure the project is on track. They will perform “spot checks” to make sure Guidelines and SOPs are being followed at all levels, for all tasks (Data Collectors and Transcribers) and that all Checklists are being filled out.
The **General Manager** should also have an organized file system for all the complete data for the project -- a **master SayMore Project** complete with **Sessions & People** and all their related files (WAVs of conversation-level speech, JPGs of person metadata forms, and TXT of transcriptions). They will also be tasked with keeping backups of this data on a local hard-drive *and* the 360 Cloud, *as well as* transferring this data to the analysts.

A good General Manager will understand the full scope of the project, and be able to task others with data collection goals and deadlines. They also have excellent organizational skills, and a very good understanding of computers. They should also be aware of all the different jobs in creating a Speech Corpus, and who is responsible for what.

1. **Familiarize yourself with each of the roles in the project by actually doing their jobs:**
   a. **Data Collector** -- Collect speech data in your native dialect; follow the SOP for Data Collectors, and record People's speech as a WAV file; collect the metadata and properly name your files
   b. **Transcriber** -- Follow the SOP for the transcriber; use Audacity to segment your WAV into conversation-level WAVs; use SayMore to enter your metadata, import the conversation WAVs to the Session, and transcribe them
2. **Review the Manual's section on Representation**: one of the most important jobs of the 4 managers together is coming up with strategic plans to make sure the corpus is representative, balanced, and well-organized; if most of the data coming in is from male speakers, for example, you will have to communicate the need for more recordings of female speakers
3. **Formulate a Plan of Action** with your team for recording speakers of all dialects, all ages, all sexes, all education levels, and all occupations, etc.
4. **Review the data that is being collected by your staff**; perform **spot-checks** by listening to some (but not all) of the WAV files and reading their transcripts to test them for quality
5. **Review the New Words List** and discuss the entries with your management team; make decisions for which spellings to make standard, and communicate that decision to the transcribers
6. **Keep your file system organized, up-to-date, and backed-up (both locally, and on the cloud)**; Keep in touch with the **Data Analysts** for data transfers and any questions that come up

This is the SOP, in brief, of the General Manager. Please develop the position with your own ideas for improving the project! Please refer to the **Guidelines** and **Checklists** for more details, and *use your common sense*. If you have questions about any of the data you are collecting, please go to the **General Manager** for their input. Also feel free to ask the outside analysts and specialists for their opinion.
Person (人员)
2 Person (人员) -- Collecting People's Speech Data

Above, we discussed the Project-level Standard Operating Procedures (SOPs)—the duties and responsibilities of the Managers. In this section, we will outline the job of the Data Collector, who is in charge of collecting people's speech in the field.

2.1 Data Collector SOP

The first job in creating a speech corpus is collecting speech data. This job requires locating potential speakers (persons, 人员) who 1) fit the sample population targeted (for example, a female student who speaks Amdo dialect; an adult, male, blue-collar worker from Lhasa; etc.) and 2) are willing to be recorded anonymously.

A good speech-data collector is comfortable speaking to a wide range of people, explaining the project in terms they’d understand, and getting their permission to record speech. They also need some working knowledge of basic recording equipment, like how to use it and how to care for it.

In addition, a good speech-data collector needs to be very organized. They need to very carefully keep track of which recordings belong to which speakers, keep the files named according to the guidelines, and record the metadata according to the guidelines.

After collecting your speech data, you will have one paper document and two digital files to give the Manager: 1) the paper version of the person metadata form; 2) the .JPEG digital photo of that metadata form; and 3) the .WAV file of the speech data. Both the JPEG and the WAV will be named to match each other; they will have the same file code, which will follow the naming conventions.

We’ll cover the details of the files later. For now, let’s discuss the recording equipment you’ll be using...

2.1.1 Recording Equipment

The first thing the speech-data collector needs is a digital recorder. Your recorder should hold a minimum of 8 GB—eight gigabytes—and be able to record in a standard .WAV file format. This is to ensure that it can record a full day's worth of high-quality audio in one go. WAV is also the file format the transcriber will need later on.
[PHOTO OF SOUND RECORDER]

Your recorder should be small, lightweight, and fit easily in a pocket or some other secure place on the person (人员) you are recording. It should also come with a clip-on microphone—a mic that clips onto a person’s collar—that clips on securely and discreetly.

[PHOTO OF SOUND RECORDER attached to person]

If you are working with multiple audio recorders for multiple speakers, be sure to label your recorders. When you fill out your metadata form, keep track of which recorder your person is using. That way, you won’t mix up which recorder goes with which recording.

If you are buying a new sound recorder for this project, make sure that it has the following:

Recorder Specifications:

1. 8 gigabytes or more of space
2. Records in standard .WAV file format
3. Comes with, or is compatible with, a clip-on mic

2.1.2 Recording Metadata

Your Manager should be able to assign you a type of person (人员) to record. At the beginning of the project, this will include just about anyone; later, there will be specific demographics that may have been overlooked. If you aren’t sure whose speech to record, please ask your manager.

Note that there is a manager for each of the 3 main dialects: Amdo, Kham, and Central Tibetan. Your dialect should be the same as your manager’s. You will be recording speakers (persons, 人员) of the same dialect. In other words, there will be 3 managers, and within each, the manager will speak the same dialect as the Data Collector (you), who will collect data on speakers (persons, 人员) who also speak the same dialect.

The type of person (人员) refers to the metadata you will be collecting on each and every speaker you record. It is important to explain that, besides this basic data, your speakers will be completely anonymous. Details like their name will not be collected, and the recording of their voice will never be made public.

The Metadata Form
The only data collected will be the very general questions of the metadata form. It is important that you, the data collector, personally speak with the person you are recording and fill the form out yourself. Do not ask your speaker to fill the form out for you. As a data collector, you are responsible for understanding how the form works, what it is for, and how to fill it out properly. The speaker has no other job but to wear the recorder, forget that they are wearing it, and then return it to you the following day.

2.1.3 Recording People’s Speech

Step 1: Prepare

First, ask your manager what kind of persons (人员) need recording. Early in the project, they will most likely answer “Anyone.” Later, they may get more specific, and say something like, “We have lots of recordings of male speakers. Please record only female speakers for now.”

Even further into the project, they may require you to collect the speech of very specific age groups, occupations, or locations. Brainstorm. Think about where you will go to find your speaker, and make a plan. Will you find them easily and nearby? Or will it require some travel? Be aware that in some cases, it may take you a day or two to find a speaker that fits the profile. If you’re really not able to find the person given by your manager, simply ask for another profile.

Next, make sure your recording equipment is ready. Plug in your sound recorder the day before going out to find speakers; make sure it charges overnight, and in the morning, test that its battery is fully charged. If you have multiple recorders, make sure they are labeled (label them Recorder #1, Recorder #2, etc.). Make sure they are set to record in WAV file format.

Finally, print out copies of the metadata form. Make sure you’ve reviewed what each section of the form means; pay especial attention to the OCCUPATION question, as it is the most difficult. (Read over the Guidelines for Person Metadata). Make sure you bring your (fully charged) phone with you to take a clear picture of the form after you’ve filled it out. When you head out the door, make sure you have these things with you:

1. Your sound recorder, set to record in WAV, with a full charge and a clip-on mic
2. Your metadata form and a pen that works (bring extras)
3. Your fully-charged phone (or other digital camera)

Step 2: Finding a Speaker (person, 人员)

Follow your plan. If you’ve been asked to find a student, for example, go to the school or to a place that students (or their parents) hang out. (Ask permission by phone or email first if
appropriate). If you have been asked to find a farmer, travel to a farming community; a nomad, the countryside; and so on.

Again, a good **data collector** needs to be comfortable speaking with a wide variety of people. When approaching people, decide where and how is best: In some cases, a phone call or WeChat will be appropriate; other times, a friendly, in-person approach might be best. Use your best judgment as to where and how to meet people.

After you find a speaker, explain the project to them using language they will understand. Be sure to tell them the basics of why you’d like to record them, and be sure to get their permission to do so. Also explain that you don’t need any personal data from them (like their name or contact information), that the voice recordings will be transcribed by someone else (who doesn’t know them), and that the voice recordings won’t be released publicly. Their only “job” is to wear the recorder, **forget that it is on**, and then return it to you the following day.

After you receive their permission to record them, you will ask them a few general questions (and immediately record their answers on the metadata form). Follow the metadata form line-by-line as you ask them the questions, and immediately record their response. It may sound silly, but if you don’t do it right away, you might forget. Be sure to **do it then and there**.

The metadata form has **only five questions**: 1) What year was the speaker born? (Ex. 1981); 2) Are they male, female, or other?; 3) What is their native dialect? (it should be the same as yours!); 4) Were they educated primarily in a school or a monastery? And what is their highest completed grade in school? (a number, 0-20+); and 5) What is their occupation? (ex. Student, Monastic, Homemaker, etc.).

You may chat with your speaker if they are open to talking more about themselves; however, beyond making plans as to where to meet to collect the recorder back, you do **not** need to record any data other than this. Immediately after filling out the metadata form, take a picture of it for a digital record. Keep both the paper and the digital copy; you will return these to the manager, along with the sound file, in order to receive payment.

Finally, clip the mic on their collar and begin the recording (in WAV format). Make sure the mic and recorder are secure. The recorder should be in a pocket or another place where it won’t fall out or become unplugged from the mic. Your recorder should also have a “lock” function, so that buttons can’t be accidentally pressed. Enable the lock.

In brief, Step 2 requires:

1. **Find a speaker** (explain the project; get their permission; teach them how to wear the recorder; make a plan of where to meet to collect the recorder)

2. **Record their metadata** on the paper form (you **must** answer each question; if you are working with multiple recorders, be sure to record the recorder number on the form)
3. **Take a digital photo** of their metadata form
4. **Start the recorder** after securing it on the person; make sure it is recording in WAV format, that the recording has started, and enable the lock function
5. **Recover the recorder** the next day (make sure the audio file is there, and is of good quality)

**Step 3: Organize & submit your files**

As we mentioned, the job of the data collector requires 3 documents: 1 paper metadata form, 1 digital photo of that form, and 1 digital sound recording of a person’s speech. You’ve carried out step 1 and step 2, which means you now have each of those documents. How do you organize them for submission?

First of all, the paper form exists as-is; you don’t need to do anything to it! If you do not have a computer, you will have to meet with your manager after each and every recording session. Otherwise, it’s very easy to mix up which sound file goes with which metadata form!

When you meet with your manager, you will give them 3 types of document (1 paper form, 1 digital photo, and 1 set of sound files). The manager will then check them for quality and accuracy, name them accordingly, and (if everything is in order), give you payment and your next assignment.

If you do have a computer, you can upload and organize your files to submit later, in batches of more than one. However, it is **crucial** that you upload digital files to your computer the day you receive them. If you wait, there is a very real danger that you will mix up which files go together.

**First, upload and rename the digital photo:**

1. Add the date (YYMMDD) and location of Recording (A, K, U).
2. What year was the speaker born? The first part of the filename is the four-digit year of birth, in Arabic numerals (1234, not ๒๓๔)
3. Are they Male, Female, or Other? The second part of the filename is the one-letter code for the speaker’s sex, in Roman characters (P, M, X)
4. What dialect do they speak? The third part of the filename is the two-letter code for dialect (A, K, U) and location
5. What is their level and type of education? The fourth part of the filename is a one-letter code (monastic or lay) followed by a two-digit number (0-20)
6. What is their occupation? The second-to-last part of the filename is a one-letter code for their occupation (UNEMPLOYED, STUDENT, OTHER BLUE COLLAR, NOMAD, FARMER, or WHITE COLLAR)
7. Finally, do they know the alphabet? The last code is for simple, basic “literacy”. 

27
Your filename for the photo should look something like this:

170428A1980MAL06LY.jpg

Next, upload and rename the digital sound file(s):

Recording in WAV has left you with a set of files 2-3 hours each. (If your device is 8 gigs, for example, you will end up with 4 WAV files of 3 hours each, or 12 hours of recording time). Name them using the same name as your photo file! You should now have two file types, a set of 1 (or more) WAV files and one JPG. The should be in the same folder, and bear the same name, since they refer to the same speaker!

170428A1980MAL06LY.jpg
170428A1980MAL06LY1.wav
170428A1980MAL06LY2.wav
170428A1980MAL06LY3.wav
170428A1980MAL06LY4.wav

You are now ready to either submit the files to your manager; or, continue collecting more speech data. **DO NOT collect more speech data until you have safely stored and renamed your files.** Your sound files will be very large (2 GB each). Make sure you have enough space to store them; having two backups, one on an external HD and on the 360 Cloud (https://eyun.360.cn/), is best.

Your manager will be checking the quality of your data when you submit it. **If you are not able to submit all 3 documents, properly stored and named, we won’t be able to pay you for your work!**

Follow the Person Checklist to make sure all your work is complete.
The job of the **data collector** ends here; next, we will describe the job of the **transcriber**!
Session (任务)
3 The Sessions (任务) -- Transcribing Speech Data

3.1 Transcriber SOP

Now that the data collector has finished collecting metadata on the person (人员) and recording their speech, it is time to start the transcription process. A good transcriber is familiar and comfortable with using computers. They will need to process audio data (cut large audio files into smaller segments) and transcribe speech using specialized software.

A good transcriber is also good with language. They are an expert in speaking and writing their own dialect (Amdo, Kham, or Central), and, while they have excellent grammar and spelling, they understand that “writing correctly” is different from “transcribing correctly.” “Transcribing correctly” means recording speech exactly as it is spoken.

If a speaker makes a grammar mistake, a good transcription contains that grammar mistake. If a speaker doesn’t finish a sentence, a good transcription contains an unfinished sentence. If a speaker speaks poorly, a good transcription is written poorly. What makes a transcription “good” is not how nicely it is written, but how closely it records exactly what the speaker says.

If a speaker uses a speech word that the transcriber doesn’t know how to spell, and they cannot find it in the dictionary, the transcriber must make note of: 1) the word; 2) the time in minutes and seconds it appears in the sound file; and 3) search for a thread on the “New Words” shared spreadsheet document (in the 360 Cloud). If the word is found, they need to make a comment on the word. If the word isn’t found, they need to create a new entry.

We’ll go into more details on these points later. With that said, the first job of the transcriber is to process the data: First, the sound file needs to be cut into manageable lengths, and all the gaps of silences and background noises (where there is no language data) need to be cut out. Each of these pre-processed sound files becomes a “session” (任务). After updating your manager with the session files, the transcriber may begin transcription.

Here is how to proceed:

3.1.1 Pre-Processing the Speech Data

First, due to the specialized software we’ll be using, it is a requirement that the Transcriber works on a Windows machine. Because of the file sizes we’ll be working with, it is also recommended that the Transcriber has a modern machine with at least 8 GBs RAM, a
decent CPU (comparable to the i5 or faster), and plenty of hard drive space (500 GBs or higher). Running a legitimate copy of Windows is also highly recommended for both speed and stability. If you are buying a machine specifically for this project, take these specification suggestions into consideration.

**Step 1: Prepare the files for pre-processing**

You should receive two files from the data collector (via the manager): 1) a JPEG of the metadata form and; 2) a set of WAV files (the speech recording). They should all have the same name, which is also a code for the metadata details of main speaker in the files (their age, sex, dialect, education, and occupation). Upload these files onto your computer. Keep them together in the same Session Folder, which also has the same name.

**Step 2: Cut the audio file into multiple files**

First, download and install Audacity: [https://sourceforge.net/projects/audacity/files/latest/download](https://sourceforge.net/projects/audacity/files/latest/download)

If your Windows platform language is set to Chinese, choose “English” -- Audacity should still automatically install in Chinese. If not, you may change the language settings in: Preferences -> Interface -> Language:

Open Audacity. Go to File -> Open... or hit CTRL+O. Navigate to the first WAV file, and select it. Audacity will prompt you with a warning, prompting you to make a copy of the WAV before
editing it. **Don’t change the default setting.** Hit “Okay,” and allow Audacity to make a backup copy. This should take a few minutes:

When Audacity has finished copying and loading the file, you should have what’s called a “waveform” view of the file. A “waveform” is a graphical view of a sound file. Quiet sections of audio are small, while loud bits of audio look big. Your screen should look something like this one; below you can see the file is quiet in the beginning, with a series of sounds happening only in the second half:

We’ll now be cutting the audio into manageable sizes. Familiarize yourself with how Audacity works. Press the **spacebar** to play the sound file. Press it again to stop. Try clicking and different sections of the sound. Listen to the conversations on the file.

Next, chunk the file by “splitting” it. For splitting, we’ll be using the **Edit -> Clip Boundaries -> Split** function: **CTRL+I**. Here, I’ve moved my cursor to just before a conversation begins. I’ve clicked my cursor once, and then hit **CTRL+I**. My sound file now has two sections:
I've double-clicked on the first section to select it (Audacity makes my selected section grey). Since this section is silence, I'll delete it by hitting my DELETE key.

Now I'll split my files into the different natural conversations. Listen to each section carefully. Make sure you don't delete anything unless it's silence or background noise:
Delete sections that are silence and/or background noises; next, we'll label each conversation according to Setting, Audience, and Communicative Aim. In a natural speech recording, how do you know where to cut?

1. SILENCES -- There is a long gap of silence (or background noises) where no speech is taking place.
   a. Cut the silent portions out; the old file ends where this cut takes place, and the new file begins where the speech picks up

2. CHANGES -- There is a change in:
   a. Setting Changes -- make the cut at a convenient pause whenever speech changes settings
      i. Ex. A conversation between two people moves from inside a private home to outside on the street
   b. Audience Changes -- make a cut at a convenient pause whenever speech changes audiences
      i. Ex. A conversation between two people is interrupted when a third person enters the house and joins in
   c. Communicative Aim Changes -- make a cut at a convenient pause whenever the purpose of the speech changes
      i. Ex. A conversation about making plans turns into an argument

In other words, to properly pre-process speech data, the transcriber needs to have a very good understanding of the session metadata: the categories of the kinds of speech data that have been collected. Just as the data collector has recorded metadata about the kinds of speakers (their age, sex, dialect, education, and occupation), the transcriber is responsible for recording metadata about the kinds of speech that speaker has engaged in.

There are 3 main categories to consider: 1) the setting; 2) the audience; and 3) the communicative aim. The setting can be: Institutional, Public, or Private; the audience can be: World/Many, Few, Single, or Self; and the communicative aim might be to: Inform, Influence, Express, Social, or Logistical. Refer to the Guidelines for Session Metadata for more details. Familiarize yourself with all the categories.

Again, start by cutting out obvious gaps between speech: long lengths of silence or background noise where no one is speaking should be cut. Then cut conversations into units following the Metadata Guidelines. Use CTRL+I for cutting and the DELETE key for deleting.

If you accidentally delete a section, don’t worry; just retrieve it using CTRL+Z (undo). If you get to a point where you’ve changed too much, making too many mistakes, don’t worry; that’s why we had Audacity make a copy at the beginning. Exit Audacity, without saving your changes, re-open, and start over.
Next, we’ll be adding labels to our audio sections. Double-click to select your first conversation. Then hit **CTRL+B** (EDIT -> Paste Text to New Label).

A **Label Track** section should appear below the waveform of your sound file. Give the label the full name of the track, and add the metadata info for the setting, audience, and communicative aim; also add a track number (in case you have duplicate sections with the same label).

For example, if your first conversation was a private conversation (G) with one other person (J) with the object of teaching them something (N) your label would be:

**01GJN**

Continue with each conversation. Label using the shortcut **CTRL+B**. Listen and label carefully; if you need to split a conversation in two again (because setting, audience, or aim changes), use the shortcut **CTRL+I**.
Continue splitting the file (CTRL+I), labeling conversations (CTRL+B) and deleting silences (DELETE) until you reach the end of the file. Go back over each conversation and make sure you’ve labeled it correctly. If you need to re-join two split sections, use CTRL+J. When you are satisfied everything is correct, go to File -> Export Multiple (CTRL+SHIFT+L):

Choose the folder to export files to (choose the same folder where you are keeping all your files for this session). Make sure “Split files based on” is set to “Labels.” “Name files” should be set to the first option, “Using Label/Track Name.” Select “Export”, and “Okay” for each file.

Be sure to save your Audacity project. It should bear the same name as your WAV file. Check your Session folder and make sure the files are there:
You’ve successfully split the audio. Exit and check your folders. Above on the left is the “Person” folder. It contains your uncut WAVs and the Audacity file, all labeled with the “Person” metadata. In your “Session” folder is a sub-folder of the same name: the “Person”s session. In that, you should have the conversation-level cut WAVs, labeled with the conversation data.

*Note: When you move on to the next chunk of uncut data (170428A1980MAL06Y2 in our example), take note of where you left off. Here, our last cut filename for the first chunk of data was numbered “05”. You’ll need to pick up where you left off to ensure unique filenames. Start your next label set with “06”!

Prepare to rename the files!

**Step 3: Double-check your file names**

You should now have a collection of small, conversation-based WAV files. They will be labeled with a three-letter code (standing for the setting, audience, and aim) along with a track number. The full-length file is now simply a backup. **However, never delete a file.**

Make sure the files use only Roman letters (AaBbCcDd, etc.) and Arabic numerals (1234567890). Do not use Tibetan, Chinese, or any other Unicode characters in the file names (this is to make sure your labels match other workers’ labels, and help software everywhere be able to read and use the files).

After this step, don’t start transcription just yet! First, send (or give) the files to the manager of your dialect; after they have been checked for accuracy and correctness, you will be given the go-ahead to start transcription.

Follow the **Session Pre-Processing Checklist** to make sure your work is complete!
3.1.2 Transcribing

First, download and install **SayMore**:  

**Step 1: Project**

When you open SayMore for the first time, it will prompt you to create a new project. Choose, “Create new, blank project.” SayMore will then ask you to name your new project. Give your project the **same name** as your person.
Hit “OK.” SayMore’s default language is English. Now that we’ve opened it, we can change it to Chinese if we want. Go to Project -> Change User Interface Language:

Select “Chinese,” and hit “OK.” Now fill out the Project Metadata in the About This Project section:

- **Title:** We will create a new project for each person; type in your “Person” code;
- **Description:** *leave blank*;
- **Vernacular:** click “Change Language;” type “Tibetan (in English characters)” into the search field; find “Tibetan; bo; 3 countries,” “Amdo”, or “Kham” and hit “enter”; or, *leave blank*
- **Location:** is *your* location;
- **Region:** Is your 1-letter region code: are you in Amdo (A), Kham (K), or Central Tibet (U)?
→ **Country:** China (find and select “China”), or leave blank
→ **Continent:** Asia (find and select “Asia”), or leave blank
→ **Contact Person:** Your name goes here (unless you prefer to remain anonymous)
→ **You may leave the final fields blank**

You've now created a SayMore project that can be used for *all* the **Sessions** you will be transcribing. When you open SayMore, it will open this project by default. Now let's add the **Person Metadata** before we start transcribing!

**Step 2: Person**

Go to “Person” and select a “New Person.” SayMore will create a new **Person** for you:
Your Person Metadata is already all contained in the file name that the Data Collector assigned. It should be the name of the JPG (photo of the Person Metadata Form) and the name of all of your WAV files for the session. It will look something like this:

170428A1980MAL06LY.jpg

In case you don’t remember, these codes stand for:

1. #### -- the Person’s year of birth
2. X -- the Person’s sex
3. XX -- the Person’s dialect
4. X## -- the Person’s education and level
5. X -- the Person’s occupation
6. XX -- the location of the Recording
7. #### -- the date of the Recording

Since we’ve been keeping careful track of this information, we don’t have to worry about losing it or where to find it. Now, we can copy it directly into the Person Metadata for SayMore.
→ **Full Name:** the full code from your file name (#######.XX.XXX.XXX.XXX.XXX.XXX)
→ **Birth Year:** number 1 from above (#####)
→ **Nickname:** leave blank
→ **Code:** leave blank
→ **Gender:** drop-down, select “Male (P)” or “Female (M)”
→ **Primary Language:** number 3 from above, the Person’s dialect
→ **Learned in:** leave blank
→ **How to Contact:** leave blank
→ **Other Languages:** leave blank
→ **Education:** number 4 from above (X##)
→ **Ethnic Group:** leave blank
→ **Primary Occupation:** number 5 from above (X)

Finally, click “Add Files” and add the JPG (the photo of the Person Metadata Form) to the Person (“Add Files…” comes with a small green “plus” sign, and can be found on the far right of the screen):

Congratulations! You have successfully added the Person to your SayMore project. You will be selecting this “Person” for all of the WAV files you will transcribe in the Sessions you’re about to create...

**Step 3: Session**

Now select “New from device…” under “Session.” Navigate to your Session folder, and select each of your conversation files (each split WAV file that you created during pre-processing). After making sure you’ve check-marked each one, click “OK.” SayMore will now create a Session for each of your conversation files.
Go ahead and fill out the details for each Session before you begin transcribing.

- **ID:** This field should be automatically filled in as the full ID of the WAV file -- it is your full code-set of metadata, like: 170428A1980MAL06LY
- **Date:** The date of recording, the last #### in the code-set
- **Title:**
- **Setting:** Your setting, audience, and aim subset: S.A.C
- **People:** Use the drop-down to choose the Person you just created; the Person should match the ID field
- **Location:** The location of the Recording, the two-letter code at the end of your file name
- **Genre:** Since we are using a different data-structure than SayMore for genre, leave the setting as-is (“Unknown”)
- **Access:** *leave blank*
- **Situation:** *leave blank*
- **Description:** *leave blank*

Now that all our metadata is properly filled out, we can finally begin transcription. In the Session, click on the WAV file. Below, choose “Start Annotating.” For **Segmentation Method**, allow SayMore to automatically segment the sound file for you (default), and click “Get Started...”
SayMore will take some time to segment your audio into small, manageable sizes for you. Once it does, the Transcription window will appear. An Annotation file has just been created for the Session. You'll see it as a sub-file of your WAV in the Session window.

The first thing to do is to set a larger font, since the default size for Tibetan is too small to read. In your Transcription window, select Options -> Fonts... and select a font size. (18 should be large enough, but experiment and see what size works best for you).
You can now play each section individually by clicking on it. SayMore will automatically loop the recording. First, make sure you are familiar with the supplementary document: **Guidelines for Session Transcription** (任务). This document gives detailed instructions on how to deal with issues of spelling, disfluent speech, and more.

Listen carefully several times as you type your Tibetan transcription into the “Transcription” section. **Make sure to type exactly what you hear.** Do not make corrections or additions or deletions to the text. After you’ve finished typing each section, listen again as you read it. Does your text match the speech word-for-word?

SayMore allows you to slow down the playback. This is an important tool to help you hear all the details of a speech recording. It’s especially useful if your speaker is speaking too quickly to be understood easily. Experiment with different % of playback speed by selecting them from the drop-down menu:

Transcribe each line. When you’ve finished with the first Session, continue to the next. SayMore will automatically save your data as you go. If you wish to change how the sound file has been segmented, you may use the “Segment...” (分段) tool:

When you’ve finished transcribing all of your Sessions, you will have a **Project Folder** in your SayMore folder (usually, SayMore puts this in your main “Documents” folder in Windows). This
folder (labeled “BO-K.R”) is what you will submit to your manager. It contains subfolders for the People & Sessions associated with the project, along with a SayMore file (extension .sprj).

Follow the Session Checklist to make sure your work is complete!
4 What Next?

4.1 History of Corpora

During the project, we will be collecting speech and transcribing it, while tracking its metadata. Where do we go from there? As we mentioned in the introduction, some of our first targets are literacy and education. These were some of the first applications in other languages like English, too. The first corpuses were small—around 1-3 million words—but still incredibly effective! As our corpus grows, so, too, do the applications it is useful for.

Let’s take a quick look at the history:

- **1953** - Publishers begin thinking about readability & plain language; Michael West publishes the first General Service List based on a 2.5 mil word corpus
- **1953** - 出版商开始思考可读性与简明用语的问题; 麦克西部根据两百五千万字的语料库出版了《基础单字字库》(General Service List)
- **1960s** - Brown Corpus is compiled, a well-balanced 1 million words
- **1960s** - 布朗语料库总集了一个相当均衡的一万字语料库
- **1980s** - COBUILD begins a 7.6m word corpus for Collins Publishing; their dictionary is published 7 years later
- **1980s** - COBUILD 开始为科林斯出版社搜集七百六十万字的语料库，该出版社在七年后根据这个语料库出版了柯林斯字典
- **1990s** - BNC (Oxford, Longman) aims for 100m; the second COBUILD hits 245m
- **1990s** - 英国国家语料库 (剑桥, 牛津) 达到一亿字; 第二版的柯林斯语料库达到两亿四千五百万字。
- **2000s** - Corpora start reaching word sizes of the billions, & are used in education, translation, dictionary creation, and more...
- **2000s** - 语料库开始达到十亿字，并且被运用在教育、翻译、字典编纂等领域里。
We can summarize by saying that corpuses were first used for literacy (Plain Language & Readability applications) and education (Graded Reading and Textbooks). These corpora were small—only 1-3 million words—but extremely effective. Beginning in the 1950s, the average grade level needed to understand the newspaper was halved. Readership and sales increased by 60-90%.

In the following decades, dictionary makers begin to realize the usefulness of corpora. COBUILD and Oxford set the standard for reference works—all modern dictionaries now take advantage of corpus data. Specialized Learner's Dictionaries are created that combine corpus-based readability for definition writing with corpus-based data for entries.

Corpus size grows steadily with advancing technology. The larger digital storage devices get, the larger corpus size can grow. It is common for corpora for General Use — Representative Corpora — to number in the hundreds of millions of words. Both COCA and BNC now weigh in at around half a billion words.

Meanwhile, Google pioneers a corpus-based, statistical approach to machine translation beginning in the early 2000s. The technique requires bilingual and monolingual corpora numbering in the hundreds of millions (for bilingual data) and billions (for monolingual data). For the most advanced applications, the more data the better.
4.1.1 What does history tell us?

We know that the more data we have, the better. Remember our fish jump example? One fish jumps, and it doesn't give us very much information. But the more fish we observe, the more useful our data became. It's the same with language. Huge corpuses allow us to observe very nuanced data. They allow us to see very rare words, in very rare circumstances. A small corpus is very unlikely to contain rare words...

But we also know that some data is better than none, and that some applications don't require lots and lots of data. The earliest and most effective ways we can use corpora are for education—at the lowest levels of language (second language learners and children), we only need the most obvious and easiest language. Even a small corpus of 1 million or more words will start being useful for these applications.

语料库的应用

- 编写教科书：根据语料分析结果，设计不同语言程度的教科书
- 语言分频读物：可显示不同语言程度词频高低的可阅读性编辑器
- 编纂词典：词义变化、常见搭配词、真实语境的例句
- 机器翻译：需要大量的语料来进行统计分析

And, as our corpus grows, so will the scope of our applications. As we get into the tens of millions, spell-checking and dictionaries become possibilities; and, in the hundreds of millions, search engines and translation look more and more plausible. But first, we'll need to
collect as much data as we can (Speech and Writing); balance it as best we can (represent all dialects, people, types, and topics of language); and post-process it to make it machine readable (Segment; POS tags; Lemma tags).

4.2 Analysis Tools

We’d also like to give you a taste of what’s possible when you have a large amount of post-processed Tibetan language data. Here, we’ll briefly introduce 4 free Corpus Tools that we can use Tibetan with today (and in the near future); the bigger and better our corpus, and the more accurate our post-processing tools, the more useful and helpful these tools will be!

4.2.1 AntConc

http://www.laurenceanthony.net/software/antconc/

AntConc is a corpus analysis tool. We have recently localized AntConc in Tibetan; the user interface is also available in English and, in the near future, we expect many other languages to be available. AntConc provides us with all the basic corpus analysis tools: 1) Word Lists (ordered by frequency); 2) Concordancing (all instances of a search term, sortable by the data that surrounds it); 3) Collocates (frequent neighbors of a search term); 4) N-grams (any number of word patterns found in text), and more. We’ll be providing you with a copy of AntConc in Tibetan to explore with!
4.2.2 Voyant

https://voyant-tools.org/

Voyant is another free corpus analysis tool. Rather than software that you have to install on your machine, Voyant is accessible over the web, via your browser. Just go to the website and upload a corpus (of plain text files in utf-8 unicode). Voyant doesn’t yet split words *quite* right for Tibetan, but this is a step we’ll be working on this year. With Voyant, you have all the same tools as AntConc: lists, concordancing, collocates, etc., but with interesting (and attractive) graphical interface options.
4.2.3 SmartCAT

https://smartcat.ai/

SmartCAT is a CAT tool for translators (CAT -- Computer Assisted Translation). They’ve recently added Tibetan as a language option at our request. In SmartCAT, we can upload parallel corpora—corpus that exist in both Tibetan and another language, like Chinese or English—and make smart tools like Glossaries and Translation Memories (TMs). Then, when we’re translating a new text, SmartCAT will automatically analyze it and make translation suggestions. Whenever a word from our Glossary appears, for example, SmartCAT will highlight it for us; if we’ve translated a similar phrase before, SmartCAT will tell us.
4.2.4 SketchEngine

https://www.sketchengine.co.uk/

Last but not least, SketchEngine is the premiere tool for dictionary makers. While SketchEngine also performs all the usual corpus analyses, we’ll first need large amounts of post-processed language data (a big, segmented, POS tagged corpus) to use SketchEngine’s “sketch engine.” Once we do, we’ll be able to make “word sketches”—profiles of words with synonyms, definitions, and other useful information extracted automatically.

4.3 Training Others

We hope you’ve enjoyed this training, along with the questions, exercises, and practicums that have come with it. If you haven’t enjoyed it, that’s okay, too—as long as you were able to learn something from it. As you go out to put this knowledge into practice, a part of your job (especially if you are a manager) may require that you train other people. We’d like to offer two things to remember when you give a training:

1. DO give questions to your trainees; make them THINK
2. DO give practice to your trainees; make them DO

Briefly put, people learn best when they have to do things themselves. And, when they do things themselves, they’ll remember things best if they have to do them again and again, making mistakes (and learning from those mistakes) along the way. In other words, we built this training in the hopes that you would THINK for yourselves, and DO for yourselves: with plenty of mistakes along the way. Remember: People will learn from what they do, not what you say.

Finally, if you have questions, concerns, or feedback that will help us improve this training, please contact the authors. Our goal is to make these resources as useful as possible, and to continue being available for the corpus creation process, from beginning to end. Communication will be key, and we’re here to help in whatever way we can!
GUIDELINES for the PROJECT Metadata (项目)

1 Types

- **SPEECH**—SPEECH is data that was spoken by a subject; the corpus contains a corresponding .wav file—a recording of the transcribed speech
- **WRITING**—WRITING is data that was written by a subject

2 Formality

- **Natural**—NATURAL speech or writing is "off the cuff" -- it is speech or writing that occurred spontaneously and naturally; it is unedited & unprepared by subject or anyone else, and thus completely informal.
- **Prompted**—PROMPTED speech or writing is given in response to a prompt. The subject responds to an article online; the subject gives a talk in front of an audience; is interviewed; etc. It is speech or writing that is unprepared, but that the subject lightly ‘edits’ in real time (in response to being self-aware of the more formal context and/or audience).
- **Scripted**—SCRIPTED speech or writing is a step more formal than NATURAL or even PROMPTED language. It is something the subject prepares, such as a speech, a script for a TV show, etc. Also included in this category is any prepared writing (such as a student essay) that is unpublished.
- **Published**—Finally, PUBLISHED speech or writing is published in a recognized source. It’s an official, published transcript of SPEECH, or an edited, published article, letter, or book by an author. It is heavily edited, most likely by multiple people; it is completely formal.
GUIDELINES for the PERSON Metadata (人员)

1 Age

- #### -- AGE is the age of the speaker given by the four-digit year of birth (the age can be determined by subtracting YOR, year of recording, by YOB, year of birth -- the YOB for a text is the year of publication);

2 Sex

- P/M/X -- MALE (P), FEMALE (M), "Other/undetermined" (X)

3 Birthplace

- A/K/U -- BIRTHPLACE is the location of the speaker's hometown / homeland / birthplace, which also represents their DIALECT (DIALECT takes precedence, so if the speaker was born in Amdo but raised in Kham, and speaks a Khampa dialect, the location is marked Kham)

4 Education

- A## -- EDUCATION is a number (years of education completed) prefixed by a letter representing whether that education was monastic or lay

5 Occupation

- UNEMPLOYED
- STUDENT
- OTHER BLUE COLLAR is other manual labor (besides nomad or farmer), like mechanic or factory worker (goods-producing jobs)
- NOMAD
- FARMER
- HOMEMAKER
- MONASTIC is a specialized pink-collar job: a monk, nun, or other professional religious worker (ie, ngagpa)
- WHITE COLLAR is professional, managerial, or other high-skilled / high-trained work
GUIDELINES for the SESSION Metadata (任务)

1 Setting

- **Institutional**—Speech or writing in an INSTITUTIONAL setting: an office, government building, school, church (Buddhist temple), hospital, etc.
- **Public**—Speech or writing given in a PUBLIC setting: a conversation in a cafe or on the street, park, or another public place; a public speech, newscast, or broadcast; a blog posted to a public forum; a comment written on a public blog post; etc.
- **Private**—Speech or writing given in a PRIVATE setting: a conversation at home, a private letter, email, or message to an individual, etc.

2 Audience

- **World / Many**—Speech or writing given to an ENUMERATE audience (an undefined, wide, general audience): all public writing; television broadcasts; etc, or a PLURAL audience means a large but defined audience that is present (or specifically pointed to by the author): a speech to a roomful of people; a letter addressed to a particular organization or group; etc.
- **Few**—INTIMATE speech or writing is for a well-defined, targeted, small group of people (3-5)
- **Single**—INDIVIDUAL speech or writing is meant for one other person (two participants total). A conversation on the phone or between friends; a letter to one recipient; etc.
- **Self**—SELF is speech or writing with no audience intended: talking to oneself, taking notes or making diary or journal entries, etc.

3 Communicative Aim

- **Inform**—INFORM is to teach, explain, give directions, or describe something; textually, this includes reference texts (grammars, dictionaries), instruction manuals, newspaper articles, and the like
- **Influence**—INFLUENCE is when a speaker or author tries to argue, persuade, compromise, or hedge; textually, includes rhetoric and other texts that make arguments for or against something
- **Imagine**—IMAGINE is to entertain, tell a joke or story, or to play a game; textually, this includes literature like fiction, novels, plays, and stories
• **Express**—EXPRESS is when a speaker shares feelings, emotions, or experiences; textually, poetry and song lyrics often fall in this category

• **Social**—SOCIAL includes all kinds of communication done for the purpose of social interaction: greetings, small talk, gossip, socializing, etc., along with cultural or religious talk; textually, ritual & religious texts are given this category

• **Logistical**—LOGISTICAL communication is to conduct business, buy/sell, bargain, making requests [to get or refuse something], or make plans; “logistical” writing would be things like business conducted by email; shopping lists, etc.
Guidelines for Session Transcription (任务)

1 Spelling

Transcribers should use standard spelling whenever possible. Always check the dictionary spelling first. Some words in some dialects might be pronounced differently from this standard spelling. For example, अेल्लो (a little bit) is spelled अेळ्लो even if it isn’t pronounced that way.

1.1 New Words

Still, transcribers will come across dialect-specific words or other speech-words that aren’t in any dictionary, and don’t have a standard spelling. Some speakers even make up their own new words. These must all be carefully documented, Take a note of the sound file and document name, the time the word occurs, and your proposed spelling.

1.2 Numbers

Numbers should always be spelled out: For example, write छानूँ not छानूँ or 1.

1.3 Abbreviations

Don’t abbreviate—always spell things out fully. For example, write अन्न not अन्न.

2 Disfluent Speech

Speakers often repeat themselves, or start a sentence, stop, and then continue on a different train of thought. It is very important for transcribers to try, as best they can, to write speech down exactly as it is spoken. This includes partial sentences, partial words, grammar mistakes, and repetitions.

2.1 Grammar Mistakes

Transcribing speech is a very different skill from being a good writer. A good transcription, exactly as possible, records what is spoken. That means that when a speaker makes a mistake, a good transcription records that mistake exactly as it occurred.
The goal of the transcriber isn’t to produce pretty, beautiful-sounding, well-edited texts for readers to read; it is to produce accurate data for specialists to analyze. And the more accurate the data is, the more useful it is. For example, if a specialist wants to analyze the differences in “how people speak” and “how people write,” transcripts that are edited by good writers become useless!

Example:

2.2 Filler Words & Interjections

Speakers often make non-word sounds to fill the space between thoughts. If you find new ones, keep it on the list. These sounds are recorded like this:

| जोर | अ | अण्डा | अण्ड्य | अण्डेण | अण्डा | रेण |

2.3 Partial Words & Partial Sentences

Speakers often make half-words, half-sentences, or half-thoughts. It is important to record these as closely as possible. If a speaker stops mid-word or mid-sentence, you may record the final word they spoke (if you understood it). A “partial thought” is represented by a row of three tshegis: ⋯.

For example:
3 Other Markup

3.1 Multiple Speakers

Label each transcription section that is not the primary speaker (the one belonging to the metadata of your file). Primary speaker sections may be unlabeled. If you cannot hear or understand the speech of the person they are talking to, do not transcribe it.

However, if the other speaker’s speech is clear and understandable, label it with the Tibetan number “ฉ”. If there is a third person clearly audible on the recording, use the Tibetan “ฉ”, and so on. (Remember, all numbers spoken by a speaker should be spelled out, so Tibetan numbers should only be used to indicate a second, third, or fourth speaker).

For example:

There should not be two speakers transcribed in the same section. If there are two speakers in the same section, use the “Segment” function in SayMore to split that section.

3.2 Unclear or Unintelligible Speech

Sometimes, the recording quality may be poor, or the speaker may mumble or say something you can’t understand. You should record your best-guess (if possible) between double parentheses "//". If you truly can’t understand it at all, leave the parentheses empty inside: //.

3.3 Other Non-Words
There are a few other non-word vocal data that transcribers should record. Record these vocal data by placing a description in a set of single parentheses. Keep the group up-to-date on the suggested standard you have chosen. They include:

<table>
<thead>
<tr>
<th>Non-Word Data</th>
<th>Transcript Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laughter</td>
<td>/ʰɑɪ̯/</td>
</tr>
<tr>
<td>Tsk-tsk-ing</td>
<td>/???'k/</td>
</tr>
</tbody>
</table>

3.4 Anonymizing the Transcript

Words that give away personal information should not appear, in context, in the transcript. Personal data and private information should remain private—and it is part of the transcriber’s job to keep such data private.

Information like phone numbers, names, and addresses should be given an anonymous substitute: For example, a specific name should be noted in parentheses as a general name /ˈjɛnɪ/ . An anonymous phone number would look like this: /ˈ09837466/ .

<table>
<thead>
<tr>
<th>Personal Data</th>
<th>Transcript Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Name (ie, ជីំែ់)</td>
<td>/ˈjɛnɪ/</td>
</tr>
<tr>
<td>Personal Phone Number (ie, 09837466)</td>
<td>/ˈ09837466/</td>
</tr>
</tbody>
</table>