Tibetan Language Readability & Literacy
The Effect of Language Community on Reading Comprehension

Introduction

This study on reading comprehension was performed in the India diaspora in the final weeks of December 2013 to the first week of January 2014. In total there were 88 respondents each tested on reading comprehension on one of 4 different texts, each an excerpt of approximately 500-600 syllables. 21 of the respondents’ birthplace & upbringing was India; 25 born in Tibet and brought up in India; and 42 whose birthplace & upbringing was Tibet. Those brought up in India from a young age averaged an upper-high school level education; those from Tibet averaged a class 9 education. The youngest respondent was 17 and the eldest 49.

One major drawback of the study is that although some cross-demographic data was collected, the sample is by no means representative of the general population. Respondents were overwhelmingly male, around 25-35 years old, and fairly well educated. “Teacher” and “Student” were the most common occupations listed. If anything, then, the population sampled was more educated than the general populace. We therefore believe the results are most likely skewed toward literacy; levels of the general population are likely lower than what is presented here. Still, some important trends worth noting stick out in the data...

Overall Results

<table>
<thead>
<tr>
<th>Statistical Category:</th>
<th>Respondent Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Level</td>
<td>10.5 (grade)</td>
</tr>
<tr>
<td>Reading Level</td>
<td>3.9 (out of 5)</td>
</tr>
<tr>
<td>Age</td>
<td>29.3 (years)</td>
</tr>
<tr>
<td>Use Level of Literary Tibetan</td>
<td>1.5 (out of 3)</td>
</tr>
<tr>
<td># of sentences/phrases not understood</td>
<td>5.7</td>
</tr>
<tr>
<td># of Unknown Vocabulary</td>
<td>5.6</td>
</tr>
<tr>
<td>Total Comprehension Errors per Text</td>
<td>11.3</td>
</tr>
<tr>
<td>% Sentences containing Unknown Syntax</td>
<td>13.79%</td>
</tr>
</tbody>
</table>
The average rate of unknown vocab for readers was approximately 2%. This is a threshold number—most readers need vocabulary coverage of at least 98% to gain adequate comprehension from a text.\(^1\) Taking this vocabulary threshold along with syntax into account,\(^2\) **40% of readers tested** fell below adequate levels of reading comprehension—meaning they had significant issues understanding the vocabulary, syntax, or both—and thus had difficulty comprehending large portions of the text (16% of it, on average).

### Analyzing Data by Birthplace and Upbringing

Birthplace and upbringing are one of the strongest indicators of Tibetan language literacy. Tibetans born and raised in Tibet are more literate than their peers. They use the language more for reading and writing: their reading levels are higher, and they comprehend more vocabulary and sentence structure. By contrast, native Tibetan speakers raised in India, despite higher education levels, tend to comprehend less Tibetan language text. They tend not to use the written language on a daily basis, if at all. They have difficulty with a significant amount of the vocabulary and syntax used in literary Tibetan:

#### Comparative color code: RED - highest - ORANGE - middle - BLUE - lowest

<table>
<thead>
<tr>
<th>Statistical Category:</th>
<th>Born in Tibet, Raised in Tibet</th>
<th>Born in Tibet, Raised in India</th>
<th>Born in India, Raised in India</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Level</td>
<td>9.2</td>
<td>12.1</td>
<td>10.8</td>
</tr>
<tr>
<td>Use of Literary T.</td>
<td>1.8</td>
<td>1.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Reading Level</td>
<td>4.4</td>
<td>4.0</td>
<td>3.1</td>
</tr>
<tr>
<td>% Unknown Vocab</td>
<td>1.0%</td>
<td>2.0%</td>
<td>3.1%</td>
</tr>
<tr>
<td>% Unknown Syntax</td>
<td>7.5%</td>
<td>15.2%</td>
<td>24.7%</td>
</tr>
<tr>
<td>Total Percent of Text Affected by</td>
<td>8.5%</td>
<td>17.2%</td>
<td>27.8%</td>
</tr>
</tbody>
</table>

\(^1\) Hu Hsueh-chao, Marcella & Paul Nation (2000). [Unknown Vocabulary Density & Reading Comprehension](#).

\(^2\) Specific vocabulary errors, unfortunately, were not recorded within full sentences that were not comprehended. Rates of unknown vocabulary are therefore most likely higher than recorded, and the combined total percentage is therefore the best indicator of the literacy level.
Below we see the distribution for the frequency of reading and writing in the Tibetan language. Analyzed by country of birth and country of upbringing, it is striking to note that zero respondents who were born and raised in India read and write more than 5 hours per week. In stark contrast, nearly a quarter of respondents born and raised in Tibet do. Likewise, a full third of Tibetans from the diaspora never use the language for reading or writing, while all of those tested from Tibet use it at least 1 hr. per week.

What is so significant here is to realize that this is self-reported data for how often the respondent reads and writes on a weekly basis, in all forums. That includes social media (Facebook, etc.), email, SMS, WeChat, all internet use (blogs, news, etc.), picking up the newspaper, a pamphlet, a book, writing in a journal, writing notes or letters, and so on and so forth. A full third of those Tibetans surveyed who were born and raised in India never use their native literary language; nearly all the rest use it an hour or less per week (an average of 8 minutes per day or less):

How often is Literary Tibetan used by native speakers?

Exploring use of written Tibetan in the diaspora...

Again, a sharp contrast is seen in reading levels of those raised inside and outside of Tibet. The majority of respondents raised inside Tibet had level 4 or 5 reading comprehension (on a 5 point scale). Those raised outside of Tibet had significantly lower levels. Those born and raised in India averaged a reading level of 3, meaning they had significant trouble with both vocabulary and syntax, which negatively affected their ability to comprehend the text. This was despite the fact that Tibetans raised in Tibet had, on average, an education level 2-3 grades lower than their Indian-raised peers:
0 - Illiterate
1 - Reader understands some words and/or syntax, but fails to comprehend the text
2 - Reader understands some parts of the text, but regularly fails to comprehend its vocabulary, syntax, or both
3 - Reader gets the gist of much of the text, but has trouble with a significant amount of the vocab, syntax, or both
4 - Reader has a decent understanding of the text, but stumbles on some vocabulary or sentence structure, or relies on context to understand some meaning
5 - Reader has a solid grasp on all the vocabulary and syntax, and comprehends the text very well

Another look at the composition of the reading levels of the study: The highest level of reading includes only 2 Indian-born Tibetans of the 88 respondents. Meanwhile, besides a single illiterate respondent, no one brought up in Tibet scored below a level 3: Level 2 was comprised of over 80% Indian-born Tibetans:
When analyzed by type of error, we see a clear correlation between reading comprehension, birthplace, and upbringing. Tibetans from Tibet are better readers in terms of both vocabulary and syntax. Reading comprehension difficulties, on average, affected more than a quarter of the text for Indian-raised Tibetans...

The Importance of Spoken Language & Exposure in Readability

Although the sample size is small, the data seems to confirm two important rules about language learning. One, exposure is key: The text on Buddhism was taken from a transcript of a teaching by His Holiness the Dalai Lama. It is the only text where Tibetans brought up in India performed on-par with their peers from Tibet. For all other texts, Tibetans brought up in India were outperformed by more than a full Reading Level on the 5-point scale (by over 20%). Second, connection to oral language is important. A transcript of the words of a teacher whom everyone in India hears again and again is much easier to understand than texts to which readers have no connection.

Analyzing Data by Education & Use Levels

There was no significant correlation between levels of education and levels of literacy. Examples of this lack of correlation could be found at both ends of the spectrum: Tibetans brought up in Tibet with very little formal education but significant comprehension of literary Tibetan contrasted by Tibetans brought up in India with college-level education but very poor
Tibetan reading skills. A scatterplot of the data demonstrates this lack of correlation:

Levels of use, however, did correspond positively to levels of reading comprehension. That is, the more the respondents tended to read & write, the higher their level of reading comprehension tended to be. Outliers of high comprehension levels with low use levels did exist: these were, without exception, readers born and raised inside Tibet:

In blue, the gradation for the use of literary Tibetan is:

0 - Reader never uses literary Tibetan
1 - Reader uses literary Tibetan less than 1 hr. per week (total reading and writing)
2 - Reader reads and/or writes daily, or averages 1-5 hrs. use per week
### Analyzing Data by Text Type

Four texts were used in the study: one excerpt from a short story, one poetry text, one transcript from a teaching by H.H. the Dalai Lama, and one newspaper article. ([View the original Tibetan texts and some extended data and analysis here](#)).

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Short Story</th>
<th>Poetry</th>
<th>Buddhist Teaching</th>
<th>Newspaper Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>1,442</td>
<td>1,638</td>
<td>1,552</td>
<td>1,514</td>
</tr>
<tr>
<td>Syllables</td>
<td>511</td>
<td>590</td>
<td>531</td>
<td>579</td>
</tr>
<tr>
<td>Words</td>
<td>329</td>
<td>340</td>
<td>287</td>
<td>245</td>
</tr>
<tr>
<td>Sentences</td>
<td>27</td>
<td>66</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>Characters per word</td>
<td>4.3</td>
<td>4.6</td>
<td>5.4</td>
<td><strong>6.2</strong></td>
</tr>
<tr>
<td>Syllables per word</td>
<td>1.5</td>
<td>1.7</td>
<td>1.8</td>
<td><strong>2.4</strong></td>
</tr>
<tr>
<td>Words per sentence</td>
<td><strong>12.2</strong></td>
<td>5.2</td>
<td>8.7</td>
<td>8.8</td>
</tr>
<tr>
<td>Readability (Average Reader’s Score)</td>
<td>3.9</td>
<td>3.57</td>
<td><strong>4.14</strong></td>
<td>3.85</td>
</tr>
</tbody>
</table>
Above, we see that the best-understood text, by far, is the text based on a transcript of an **oral** Buddhist teaching—again demonstrating the connection between **oral** and **literary** language. The most difficult text, on the other hand, is the poetry.

If we apply English-based readability formulas to the texts, there is no correlation between actual level of difficulty and their readability scores—the Buddhist text, the most readable test according to the study, scores quite difficult at grade level 12.6. Meanwhile, the most difficult-to-read text, the poetry, scores the lowest in these English-based readability formulas at 6.78:

<table>
<thead>
<tr>
<th></th>
<th>Short Story</th>
<th>Poetry</th>
<th>Buddhist Teaching</th>
<th>Newspaper Article</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Readability</strong></td>
<td>7.055</td>
<td>6.78</td>
<td>12.6</td>
<td>17.15</td>
</tr>
<tr>
<td>(Coleman–Liau Index)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Why We Can’t Apply English Readability Formulas to Tibetan

Readability formulas function in English by combining two measures of complexity—complexity of vocabulary and complexity of syntax. These are estimated by taking averages of word length (measured by the average number of characters or syllables per word as an indicator of vocabulary complexity) and sentence length (measured by words per sentence as an indicator of syntactic complexity). This works in English since there is some correlation between how long words are and how difficult they are, as well as how long
sentences are and how difficult they are. Though there isn’t a one-to-one correspondence, it is true enough that, on average (over the length of the text), a sense of the overall readability can be determined.

While sentence length for Tibetan may work as an indicator for complexity of syntax, there seems to be no particular relationship between word length and word difficulty. Taking our short story as an example of this phenomenon, we see that longer sentences did result in lower levels of comprehension. The longest sentence in that text compiled a massive 45 reading comprehension errors—an average of two errors per person!

However, out of the 17 most difficult words in the text (words that 20-50% of readers didn’t know), just about all of them were a mere one or two syllables in length, some no more than three characters total. This list of commonly unknown or mis-comprehended words included: ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་ད་གཟོད་

It is on the basis of such lists that another type of formula may be a useful model for readability assessment in the Tibetan language—formulas, like the Dale-Chall readability formula,3 based on word frequency and/or difficulty.4 This assessment tool utilizes a ratio of “difficult words” per sentence in place of average word length. A list of 3,000 words that are easily comprehended by the average fourth-grader is used as the standard against which texts are measured. If a similar formula were created for Tibetan, it would serve as an extremely useful guide in the both the creation and assessment of level-appropriate texts for Tibetan language education.

A Hypothetical Readability Formula for the Tibetan Language

We may create a Dale-Chall-type readability formula for the Tibetan short story as a purely hypothetical exercise to demonstrate how such a formula could function. First, we’ll create a list of all the “difficult vocabulary” (here we’ll measure difficulty of the words by counting the number of readers who didn’t know them, from “easiest” to “hardest”; a true formula would be based from a comprehensive list culled from a much larger database):

1. བཤེར་།
2. སྣྣོ་ལོ།
3. འཞིགས་དམོད།
4. བུམ་བ།

---

4 “Words that are frequently used are usually easier than those that are hardly used. The high frequency words are often easier than the low frequency words. Hence, it is also important to differentiate the frequency of usage in determining a text’s level of difficulty.” Daud, Nuraihan Mat, et al. “A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty.” World Applied Sciences Journal 21: 168-173, 2013.
For our purposes here we will simply consider all instances of more than one reader not knowing a word as an instance of a “difficult” word (since one instance is an outlier, though adding these two extra words only changes the final result by ~0.1 [from 5.7 to 5.8]). Applying the Dale-Chall formula, we find 30 instances of difficult words. The formula is:

\[
0.1579 \times \left( \frac{\text{difficult words}}{\text{words}} \right) \times 100 + 0.0496 \times \left( \frac{\text{words}}{\text{sentence}} \right)
\]

*If the percentage of difficult words is above 5%, then add 3.6365 to the raw score to get the adjusted score, otherwise the adjusted score is equal to the raw score.*

Applying it to our short story text we get:

\[
0.1579 \times \left( \frac{30}{329} \right) \times 100 + 0.0496 \times \left( \frac{329}{27} \right) = 1.439817629 + 0.604385185 = 2.044202814 \quad \text{(and since the percentage of difficult words is above 5%, we add 3.6365)} = 5.680702814
\]

—in other words, a text that ought to be easily understood by an average 5th or 6th-grade student.

This, of course, is a purely theoretical application of nothing but a hypothetical readability formula for the Tibetan language. To compile a list of “easy” words that truly reflects the natural reading levels of Tibetans, a much broader study (such as a Spoken Corpus) is needed. Once such a list existed, however, it would be a fairly simple task to create a text analysis program to analyze the readability level of any Tibetan text.
Notes on Methodology:

Data Collection:

Text excerpts were given to native speakers of the Tibetan language; they were surveyed for their basic biographical and literary language use data. After reading the text, a trained interviewer quizzed them orally, sentence by sentence, on the text, focusing on the meaning of specific semantic structures and vocabulary. Unknown vocabulary words were circled, while confusion regarding a specific sentence was marked by underlining it (above, where the percent of unknown syntax is reported, it indicates the percentage of the text that was underlined, i.e., the percentage of sentences which were misunderstood due to semantic complexity).

Of course, this methodology opens itself up to some degree of human error; however, we believe that the results speak for themselves in how clear the patterns and correlations are. We propose further, well-designed (more objective and randomized) studies of this nature to gain a clearer picture of these trends. It would be especially useful to identify which semantic structures and vocabulary diaspora-born Tibetans have difficulty with; elementary and remedial literacy programs could then target those problem areas with a higher degree of specificity.

Word Count:

To decide what counted as a Tibetan “word” for the study, we followed a system based on English word counts. Thus, all grammatical particles were counted as “words” (i.e., “ན” and “ལ” counted as words, just as English “at” or “in” counts as a word) except for the connective particles (i.e., “མི”, which weren’t counted just as English possessives, i.e. “I’ve” and “his”, and other contractions, aren’t counted). Although this is arbitrary, and an argument could be made to count either all or none of the grammatical particles as “words,” this standard was applied consistently for each of the texts.

This not only served the purpose of simply having comparative data across texts, it also gave us the ability to compare across languages. Since there are no tools for assessing the complexity or readability of a given Tibetan text, comparisons were made with English text levels of comparable length, words-per-sentence, and other similar countable data. As mentioned above, it is shown that word-length is one significant variable that does not “translate” between English and Tibetan textual assessment.

Reading Comprehension Grades:

A grade based on adding the percent unknown vocabulary plus the percentage
uncomprehended syntax; this sum was subtracted from a total possible 100%. A reader who comprehended all vocabulary and syntax of a text thus scored 100%. Since specific vocabulary was not recorded within full sentences that were not comprehended, rates of unknown vocabulary are most likely higher than recorded, and rates of unknown syntax slightly lower (that is, when a reader didn’t understand the sentence, it was underlined; the specific semantic structure that wasn’t understood was not recorded). Therefore the total percentage is the best indicator of the literacy level.

The combined reading comprehension grade averages this “objective” score (unknown vocabulary plus unknown syntax) was averaged with the subjective score the reader received from the grader (graders gave readers a score on a 3-point scale: low, middle, or high comprehension). This combined comprehension grade places each respondent on the following 5-point scale:

0 - Illiterate
1 - Reader understands some words and/or syntax, but fails to comprehend the text
2 - Reader understands some parts of the text, but regularly fails to comprehend its vocabulary, syntax, or both
3 - Reader gets the gist of much of the text, but has trouble with a significant amount of the vocab, syntax, or both
4 - Reader has a decent understanding of the text, but stumbles on some vocabulary or sentence structure, and/or relies on context to understand meaning
5 - Reader has a solid grasp on all the vocabulary and syntax, and comprehends the text very well

Upbringing:

The cutoff age was 16—those who moved to India before adulthood were counted as brought up in India, and those who moved to India after age 16 were counted as brought up in Tibet.

Use of Literary Tibetan:

The scale used for the use of literary Tibetan is as follows:

0 - Reader does not use literary Tibetan in everyday life.
1 - Reader uses literary Tibetan less than 1 hr. per week (total reading and writing)
2 - Reader reads or writes daily, or averages 1-5 hrs. use per week
3 - Reader reads or writes over 5 hrs. per week