himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

Himalayan Linguistics

Practical applications for corpora: The role of research-based linguistics in literacy and education for the Tibetan language

Dirk Schmidt

Esukhia

ABSTRACT

Corpus Linguistics and NLP have many obvious applications for researchers, academics, and other specialists; what should not be overlooked, however, is their role in improving the mundane, everyday interactions between people and language, be they a reader of a newspaper; a child with a storybook; or a student in a classroom. The language analyses that these linguistic tools provide have an important part to play in the feedback loop between authors, journalists, and pedagogues on the one hand and their audiences and students on the other.

While these sorts of research-based resources have already made splashes in majority languages like English, their ripples have yet to spill over into the smaller language markets. Within this paper we outline the ways in which corpus linguistics may inform Tibetan language literacy and education in both L1 and L2 contexts, while drawing from our own research into issues of readability and the development of a modern pedagogy for instruction in the Tibetan alphabet based on frequency data.

KEYWORDS

Tibetan, literacy, readability, pedagogy

This is a contribution from *Himalayan Linguistics, Vol. 15(1): 167–185.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

Practical Applications for Corpora: The role of research-based linguistics in literacy and education for the Tibetan language

Dirk Schmidt
Esukhia

1 Introduction

Corpus Linguistics and NLP have many obvious applications for researchers, academics, and other specialists; what should not be overlooked, however, is their role in improving the mundane, everyday interactions between people and language, be they a reader of a newspaper; a child with a storybook; or a student in a classroom. The language analyses that these linguistic tools provide have an important part to play in the feedback loop between authors, journalists, and pedagogues on the one hand and their audiences and students on the other. While these sorts of research-based resources have already made splashes in majority languages like English,[1] their ripples have yet to spill over into the smaller language markets. There's no reason, however, that corpus linguistics should not be poised to have a dramatic impact on the way we think about Tibetan language literacy and education in both L1 and L2 contexts as we move forward.[2]

If we look back, language education was traditionally an elite skill reserved for an elite caste (European education, for example, took place in Latin until the late 17th century); without a culture of widespread literacy, reading and writing were primarily reserved for academic, official, or religious functions. A multitude of factors, both social and technological, were responsible for the seachange which pulled the written word from the formal, rarified world of clerics, lawyers, and scribes and infused our common, everyday existence with its ubiquity. Those working toward literacy and language education for Tibetan ought to be acutely aware of the history of reading and writing, its relationship to speech communities, as well as the ever-widening scope of reading and writing in daily life (the rise of vernacular literature). And, in addition to these general trends, we may specifically point to several research-based innovations that have acted as feedback mechanisms to facilitate the widespread literacy that is a mark of modernity: among these were frequency lists culled from corpus data.

[1] See, for example: the Cambridge Corpus (http://goo.gl/2yeudh) used for English pedagogy; the Collins corpus-based COBUILD dictionary (http://goo.gl/uP3WFN); or the many publishers of graded readers based on high-frequency word lists.

[2] For those unfamiliar with the terminology, L1 and L2 = First and second language. This shorthand will be used throughout the article.

What I present below, therefore, is neither technical nor innovative; instead, it is a discussion of the simple things, like frequency lists, that corpus linguistics can provide non-linguists, and how these sorts of tools can impact our view of literacy and language education for the Tibetan language specifically. Within, I ask: What if students progressed from simple, low-level texts at a graduated pace, rather than tackling sophisticated literature from the get-go? What if we could grade newspapers, novels, and children's literature to suit various reading levels? What if readability, level assessment, and graded curriculums were all part of the standard Tibetan language education? In other words, what if literacy in Tibetan was no longer an "elite achievement," as Beyer (1992) puts it, but an ordinary occurrence? These questions conclude with the theory of applied corpus linguistics put into practice: an Alphabet Book for the Tibetan language exhibiting a pedagogy based on frequency data.[3]

2 Literacy

As noted above, modernity has seen a marked increase in literacy for the general population; what remains notable about this phenomenon cross-historically, however, is that percentage of the population reading and writing at sophisticated levels has yet to exceed 20%.[4] In other words, the increase in literacy over the last few centuries has been more a matter of the written word being recast in historically unfamiliar roles, such as informal communication;[5]popular literature;[6] religious texts in the vernacular;[7] and general interest news.[8] It's no coincidence that these inroads of literacy into people's daily lives has brought about an increasingly informal literature. Writing is, after all, fundamentally a system of encoding phonemes into graphemes and decoding them back out again—and thus, the more conversational or speech-based a text is, the more easily people can read it.[9] And, as literacy has become less of a specialized skill, written texts have naturally come to reflect less specialized material. Tibetan, in contradistinction to this trend, has preserved its traditional literary form and its specialized literature; the resulting gap between the formal and informal registers is thus an issue for both L1 and L2 learners.

2.1 Diglossia

This gap between spoken and literary Tibetan is a source of much consternation for many students of the language, both L1 and L2. While both L1 and L2 educational contexts have struggled with exactly how to address this gap, linguistics provides us with a framework for understanding the phenomenon: if we assess the situation by means of cross-linguistic parallels (especially Arabic), we may suggest that the Tibetan language is a quintessential diglossia. That is, the spoken, vernacular dialects ("L", the "low" languages) are superposed by the formal, literary language ("H", the "high").

[3] The Tibetan Alphabet Book, "so ri me bu", is downloadable for free here: https://goo.gl/J67aj2

[4] Modern American high literacy (10th grade and above) is 20% (see: http://nces.ed.gov/pubs93/93275.pdf); Ancient Roman high literacy would be on the low end of an estimated at 5-30% total literacy (https://goo.gl/3mns6E); up until the modern era, literacy in China and Asia was similarly no more than 20% **total** until the modern era (http://goo.gl/xs7CZT).

[5] See: McWhorter (2013).

[6] The most popular novels are written at the 7th-grade level; see: Klare (1954).

[7] For example, see: Marsden (2011).

[8] Most news tops out around 9th grade level; most sources are easier than that, see: DuBay (2006).

[9] See, for example, the BBC Guidelines on conversational English in journalism: Allen (2003): http://goo.gl/GVdGrz.

Furthermore, Tibetan ticks all the usual diglossia checkboxes (as defined by Ferguson [1959: 325-337]):

A. there is a large body of **culturally defining literature** (in this case, the Tibetan Buddhist canon—for Arabic, the Quran);
B. the literature has **been around for centuries** (true of both); and
C. there are **low literacy rates** (an issue for both Tibetan *and* Arabic speaking populations)[10]

We may draw a causal arrow from (A) to (B) to (C): the literature of a diglossia lasts for centuries due, at least in part, to the fact that it is culturally defining; and, as the spoken language changes naturally over time, it drifts further and further from this standard, leading to low literacy. To understand why low literacy rates are an implication for diglossias also requires knowing a bit about how beginning readers learn to read. To state it briefly, they start by making associations between the speech words that they know and the printed form of those words—by internalizing the patterns and relationships between phonemes and graphemes.[11] The more explicit these connections are, the easier it is to learn to read, which is why so much focus on early reading is phonetic (the process of connecting *sounds* to *symbols*). Meanwhile, early readers in many other languages utilize language that is nearly 100% the vocabulary level of its readers.[12] Researchers have found that learning to read is challenging in and of itself, as unknown vocabulary or syntax create unnecessary difficulties for beginning readers.

In other words, it is best if early reading materials reflect the natural language levels of early readers (a.k.a. children). Their content should mirror their vocabulary; their syntax; and even their sense of humor and the ways they use language to express their inner world. In short, this material needs to be easy and relatable.[13] Yet even expert readers have been shown to be fairly bad judges of a text's difficulty;[14] this work is better left to the hard data only research can provide. The potential solid ground that early reading materials for the Tibetan language may stand on is therefore best left to corpus linguistics[15]—in this case, a spoken corpus of native speech by Tibetan children of various regions and age brackets would make for an ideal foundation. This information can then be used to "fill in the gap" between the base-level everyday, spoken vernacular and the heights of sophisticated literature. More specifically, the "filling in" process can be aided by developing Tibetan-specific readability formulas based on frequency lists culled from age-sorted spoken corpus data.

2.2 Readability

As aforementioned, nothing in this article is particularly revolutionary or unprecedented. The idea for analyzing the readability of texts using frequency lists dates back to the 1940s and 50s, when American media outlets discovered that their newspapers were too difficult for the average reader.

[10] For Tibetan, the "Human Development Report: China." *U.N. 2008*, page 140 (http://goo.gl/ovCjSK), cites illiteracy in the T.A.R. at 45.7%. If anything, this probably underestimates the rate of illiteracy since UN data is often merely based on self-reporting or the ability to write one's own name. For Arabic, see Maamouri (1998).
[11] Waring (2003).
[12] Wan-a-rom (2008: 43–69): http://goo.gl/bBw2Ox.
[13] Nikolajeva (2014). Pressley (2006). Callander (2011).
[14] Hamilton (2003: 228-240).
[15] McEnery: http://goo.gl/D6SA0A.

Lostutter's 1947 study, for example, showed that American newspapers were written at a level five years above the ability of average adult readers. Driven by a desire to boost sales, they led the charge in readability research in order to grade their own material. Within a few years time, the difficulty and grade level of the average newspaper article was halved. They accomplished this by assessing text using simple formulas to provide feedback for their writers. Among others, Robert Gunning's "Gunning FOG Index" measures the complexity of text like this (where a "complex word" is a word with 3 or more syllables):

$$0.4\left(\left\{\frac{\text{words}}{\text{sentences}}\right\}+100\left\{\frac{\text{complex words}}{\text{words}}\right\}\right)$$

The results of making texts easier to read are striking. Murphy (1947) showed that reducing the difficulty of the text in newspaper articles from the 9th to the 5th grade level increased readership 45-60%. And Swanson (1948: 339–343) found that better readability increases the total number of paragraphs read by 93% and the number of readers reading every paragraph by 82%. Modern day popular literature and news articles generally grade out at around the 7th-9th grade reading levels.[16] Somewhat paradoxically, making reading easier doesn't only mean higher literacy rates across low levels; it also functions as a net positive for higher reading levels in that the more extensively people read, the better they get at reading, and the more they read again.

Part of the reason tools like readability formulas are so important for readers and writers is that easy-to-read material *isn't* easy to write; creating highly readable texts is a learned skill that requires feedback built into the writing and editing process.[17] Authors and journalists must *train* in order to write in a simple, clear, conversational style in order to ensure that what they write resonates with their audience.[18] Using plain language across everyday domains is vital to the health and vitality of a language: as the world becomes increasingly global, and speech communities increasingly bilingual, it's more important than ever to provide people with easy, convenient options for using their mother tongue in a variety of daily contexts. But in order to apply readability formulas to Tibetan, we'll need a way to gauge the relative complexity of its vocabulary.

2.2.1 A readability formula for Tibetan

Due to a happy accident, many early readability formulas for the English language based themselves on the fact that "difficult" English words tend to be longer than "easy" ones (the former usually French or Latin derived, the latter Germanic or Norse); while imperfectly true (not all easy words are short and not all hard ones long), it is *true enough* to be useful as a proxy for word difficulty.[19]

[16] DuBay (2006). Klare (1954).

[17] See Flesch's chapter on his formula for writing Plain English (http://goo.gl/eOAp9p) and Beaglehole (2010): http://goo.gl/wt2aOl.

[18] See again BBC's style guide: Allen (2003): http://goo.gl/GVdGrz.

[19] See, for example, the Coleman-Liau index: Coleman (1975: 283–284).

This doesn't hold for many other languages, and it also isn't true of Tibetan. For example, in a small study conducted with Tibetan native speakers who were asked to read an excerpt from a short story, out of the most difficult words in the text (words that 20-50% of readers didn't know), the vast majority were a mere one or two syllables in length, some no more than three characters total.[20]

That being the case, how can we measure word complexity in the Tibetan language context? One sure way is by assigning words a value based on their frequency of use: as Daud (2013: 168-173) notes, more frequent words are more familiar, and thus less "difficult." Furthermore, words understood by younger readers are less difficult than those only understood by older readers. This idea is similar to another method used in some readability tests,[21] namely the Dale-Chall Formula, which is based on a list of 3,000 words easily understood by the average fourth grader. Again, this is where the importance of corpus linguistics can provide frequency data crucial to practical applications for assessing the difficulty of texts.

$$0.1579\left(\frac{\text{difficult words}}{\text{words}}\times 100\right)+0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

The Dale-Chall Formula measures readability by adding the average sentence complexity to the average vocabulary complexity, and then weighting their relative importance (by multiplying by a coefficient) in order that the raw score adheres to a standardized scale (0-10 in the case of Dale-Chall).

For example, a readability formula for Tibetan might be based on a measure of sentence complexity by counting the number of ཚེག་s (*tsheg*) per ཤད་ (*shad*) (the number of syllable breaks per syntactic break), or the number of words per same, and a measure of vocabulary complexity based on their corpus-determined frequency. Although we are minus a corpus, we may analyze the short story Tibetan literacy data from above and plug it into a readability formula to see what happens. In this case, 23 native speakers read an excerpt from a short story; they were quizzed orally to gauge their reading comprehension, with each instance of unknown vocabulary being recorded. If we were to apply the Dale-Chall formula to this data, there are 30 instances of difficult words in the story, which yields a Dale-Chall rating of 5.68—a text that ought to be easily understood by an average 5th or 6th-grade student.[22]

One may rightly question whether we may simply apply such a formula, unrevised, to the Tibetan language. In the case presented here, I would like to stress that this has been done as a purely hypothetical exercise, not as a tool to be widely applied as-is; instead, it is merely meant to

[20] From unpublished research we conducted in 2013. The full details of the study are available here: https://goo.gl/RknVeD. The text of the story used is available here: https://goo.gl/o82vjz This list of commonly unknown or mis-comprehended words included: བུད། གཙོམ་ཞེང་། ཞོར། ཚང་གྲོང་། བརྡལ། ཟླ་བརྒྱུད། ཡུར་བ། ཐིགས་དགོད། ལས་བྱ། ཨ་ཕྱི། ལྷ་བ། ད་གཟོད། and རྒྱ་སྲང་།.

[21] "[The Dale-Chall] formula is not the most popular although it has always been and is still considered more accurate than other formulas like the Flesch Reading Ease score, or the Flesch-Kincaid Grade Level." RFP Evaluation Center (2013): http://goo.gl/VdhWpq.

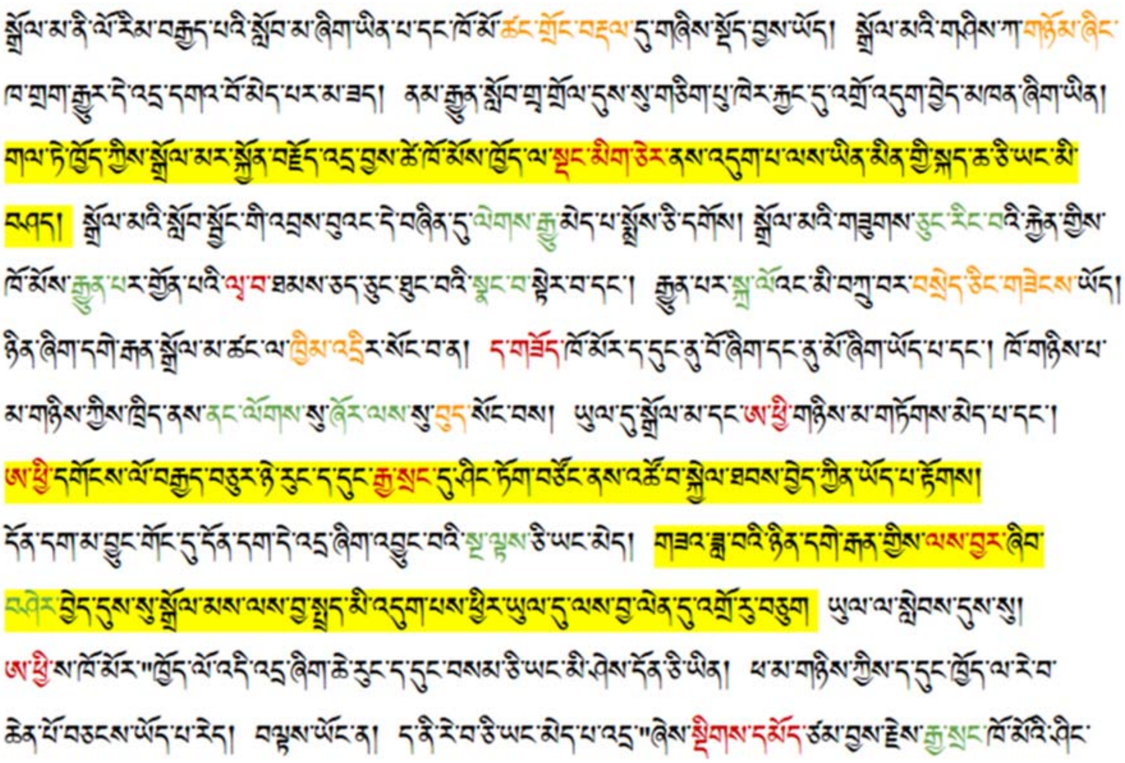
[22] 0.1579 x ([difficult words/total words] x 100) + 0.0496 (words/sentence) -> If the percentage of difficult words is above 5%, then add 3.6365 to the raw score to get the adjusted score, otherwise the adjusted score is equal to the raw score. ->0.1579 x ([18/329] x 100) + 0.0496 (329/27) = 1.439817629 + 0.604385185 = 2.044202814 -> Since the percentage of difficult words is above 5%, we add 3.6365 yielding a Dale-Chall rating of 5.68.

demonstrate the *proof of concept* that such a formula is both possible *and* useful. Issues of word-splitting, defining what is or isn't a "complex word" for Tibetan, or how to weight semantic and syntactic complexity within a purely Tibetan readability formula lie beyond the scope of this paper (and beyond the scope of any research done thus far in Tibetan language readability). However, the reasoning behind such a formula being cross-linguistically valid is sound: readers of all languages take in meaning in syntactic "chunks," and shorter chunks make for quicker understanding, while longer ones increase mental fatigue (in Tibetan, syntactic breaks are marked by the ཤད་ [*shad*], which easily machine-countable).[23] And readers of all languages tend to be more familiar with some words than others[24]—with a corpus-based frequency list and a word-splitter, vocabulary complexity, too, is a measurable quantity.

Though simple, the utility of readability formulas is far-reaching: software tools may be developed from the basis of corpus-based frequency that have the capability to show exactly which words and which sentences in any given text are likely to create difficulties for their intended audience. This would allow writers, educationalists, and journalists to edit both old and new texts to target specific age groups or reading levels. Furthermore, since a reader's ability to comprehend text is based on experienced, lived language, spoken corpus data can assist writers in using words that are phonetically charged and ripe with meaning—while providing a natural bridge from spoken to literary forms. Applying corpus linguistics to literacy issues would allow writers and educators to outline a clear path toward literacy in the Tibetan language. Below is a mockup of such a readability tool (based on our short-story literacy study from above): within, difficult words are graded by level and distinguished by color, and difficult syntax is highlighted:

[23] Flesch.

[24] See again: Wan-a-rom (2008: 43–69): http://goo.gl/bBw2Ox.



སྒྲོལ་མ་ནི་ལོ་རིམ་བརྒྱད་པའི་སློབ་མ་ཞིག་ཡིན་པ་དང་ཁོ་མོ་ཚང་གྲོང་བརྡལ་དུ་གཞིས་སྡོད་བྱས་ཡོད། སྒྲོལ་མའི་གཞིས་ཀ་གཉོམ་ཞིང་ཁ་གྲུག་རྒྱུར་དེ་འདྲ་དཀའ་པོ་མེད་པར་མ་ཟད། ནམ་རྒྱུན་སློབ་གྲྭ་གྲོལ་དུས་སུ་གཅིག་པུ་ཁེར་རྐྱང་དུ་འགྲོ་འདུག་བྱེད་མཁན་ཞིག་ཡིན། གལ་ཏེ་ཁྱོད་ཀྱིས་སྒྲོལ་མར་སྐྱོན་བརྗོད་འདྲ་བྱས་ཚེ་ཁོ་མོས་ཁྱོད་ལ་སྨང་མིག་ཅེར་ནས་འདུག་པ་ལས་ཡིན་མིན་གྱི་སྐད་ཆ་ཅི་ཡང་མི་བཤད། སྒྲོལ་མའི་སློབ་སྦྱོང་གི་འབྲས་བུའང་དེ་བཞིན་དུ་ལེགས་རྒྱུ་མེད་པ་སློས་ཅི་དགོས། སྒྲོལ་མའི་གཟུགས་ཚུང་རིང་བའི་རྐྱེན་གྱིས་ཁོ་མོས་རྒྱུན་པར་གྱོན་པའི་ལྭ་བ་ཐམས་ཅད་ཚུང་ཐུང་བའི་སྣང་བ་སྟེར་བ་དང་། རྒྱུན་པར་སྐྲ་ལོའང་མི་བཀྲུ་བར་བསྒྲིད་ཅིང་གཟེངས་ཡོད། ཉིན་ཞིག་དགེ་རྒན་སློབ་མ་ཚང་ལ་ཁྱིམ་འདྲིར་སོང་བ་ན། ད་གཟོད་ཁོ་མོར་ད་དུང་ནུ་བོ་ཞིག་དང་ནུ་མོ་ཞིག་ཡོད་པ་དང་། ཁོ་གཉིས་པ་མ་གཉིས་ཀྱིས་ཁྲིད་ནས་ནང་ལོགས་སུ་ཞོར་ལམ་སུ་བུད་སོང་བས། ཡུལ་དུ་སྒྲོལ་མ་དང་ཨ་ཕྱི་གཉིས་མ་གཏོགས་མེད་པ་དང་། ཨ་ཕྱི་དགོངས་ལོ་བརྒྱད་བཅུར་ཉེ་རུང་ད་དུང་རྒྱ་སྐད་དུ་ཤིང་ཏོག་བཙོང་ནས་འཚོ་བ་སྐྱེལ་ཐབས་བྱེད་ཀྱིན་ཡོད་པ་རྟོགས། དོན་དག་མ་བྱུང་གོང་དུ་དོན་དག་དེ་འདྲ་ཞིག་འབྱུང་བའི་སྔ་ལྟས་ཅི་ཡང་མེད། གཟའ་ཟླ་བའི་ཉིན་དགེ་རྒན་གྱིས་ལས་བྱར་ཞིབ་བཤེར་བྱེད་དུས་སུ་སྒྲོལ་མས་ལས་བྱ་སྤྲད་མི་འདུག་པས་ཕྱིར་ཡུལ་དུ་ལས་བྱ་ལེན་དུ་འགྲོ་རུ་བཅུག ཡུལ་ལ་སླེབས་དུས་སུ། ཨ་ཕྱིས་ཁོ་མོར"ཁྱོད་ལོ་འདི་འདྲ་ཞིག་ཆེ་རུང་ད་དུང་བསམ་ཅི་ཡང་མི་ཤེས་དོན་ཅི་ཡིན། ཕ་མ་གཉིས་ཀྱིས་ད་དུང་ཁྱོད་ལ་རེ་བ་ཆེན་པོ་བཅངས་ཡོད་པ་རེད། བལྟས་ཡོང་ན། དེ་ནི་རེ་བ་ཅི་ཡང་མེད་པ་འདྲ"ཞེས་སྨིགས་དམོད་ཙམ་བྱས་རྗེས་རྒྱ་སྐད་ཁོ་མོའི་ཤིང་

*A quick mockup of a readability tool distinguishing between **easy, daily vocabulary** (normal black text); **slightly difficult** words (green); **difficult** (orange); and **very difficult** (red). Thesaurus tools could help authors choose reader-friendly synonyms, and software could be set to edit for various reading levels.*

3 Language education

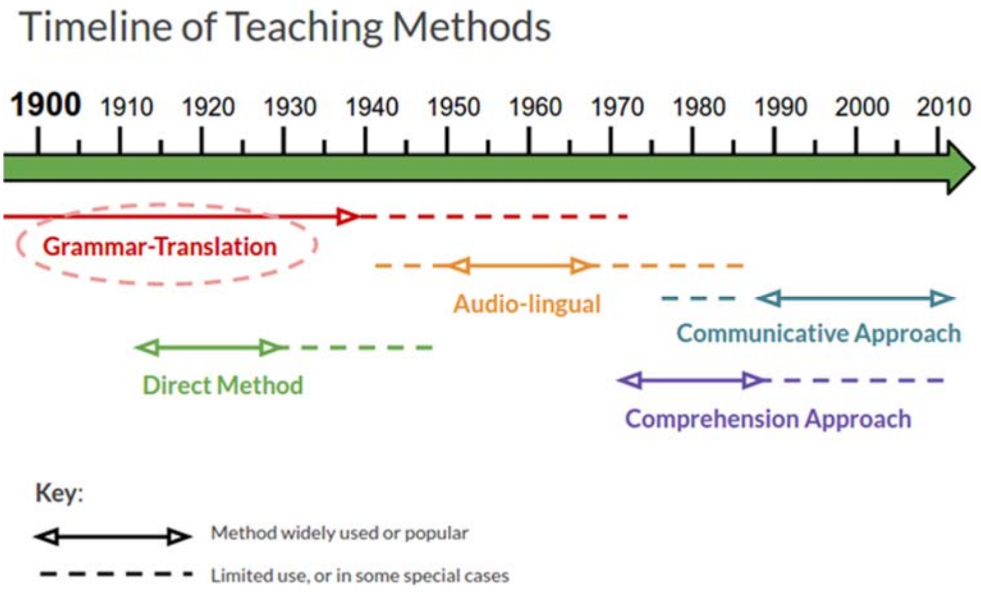
In order to fully appreciate why corpus linguistics is so important for L2 pedagogy, I think it's necessary give a brief on the current state of affairs; while I leave the heavy-lifting in this section to the footnotes and citations (and my previous papers, see Schmidt (2014)), I'd summarize by saying I believe that approaching Tibetan through the spoken language would be the *most* efficacious way for the *highest number* of students to gain Tibetan language proficiency. To do so, our primary concern is bridging the gap of diglossia. And, as it's been said of Arabic, another diglossic language:

> The field of teaching Arabic as a foreign language has benefited from the advances in foreign language teaching such as moving away from the grammar and vocabulary focused methods toward more communicative techniques (Al-Mamari [2011]).

Yet despite the progress made in interdisciplinary fields, especially those in L2 pedagogy and applied linguistics, beyond a handful of exceptions, the general method for teaching the Tibetan language remains firmly entrenched in its classical education roots and 17th century pedagogy.[25] Most students of Tibetan will be familiar with the methods of Grammar-Translation, even if they

[25] Richards (2001: 5); Huckin, (1997: 5-6) (http://goo.gl/eFWPl); and Harmer (2007: 48-49).

haven't heard the term before: studying in an English language medium (or another L1), students memorize vocabulary and study texts by parsing them grammatically, word-by-word.[26]



Grammar-Translation has its roots in the classical education of ancient Rome and the middle ages (from circa 5th century), but is vestigial in the sense that the European vernaculars replace Latin as the medium of instruction in the late 17th century (Tunberg [2012])... As Latin is no longer used for communicative purposes, but merely ***described*** *within European vernacular contexts, Grammar-Translation is born (Timeline from Taylor [2003]).*

While this method addresses the end aim of most Tibetan studies students (textual translation and analysis), by restricting itself to grammar and vocabulary, it actually lacks the ability to impart real language skills to its students[27]—as Hillocks (1991) sums up the linguistic evidence on the matter, "Research over a period of nearly 90 years has consistently shown that the teaching of grammar has little or no effect on students." Indeed, rather than bridging the gap left by diglossia, Grammar-Translation exacerbates it, while ignoring great strides in the field of second language education. (Japan's English L2 methodology is a prime example of how Grammar-Translation fails its students in comparison with other, more effective pedagogies).[28] As it's been summarized by experts in second language learning:

[26] Rockwell (1991), for example, admits in the preface to his primer that "the fundamental approach of [his] text is descriptive" and based on sentences removed from a larger context. Beyer similarly asserts on the first page of his introduction that his work is descriptive in nature, and explicitly states it is not his intention to address language *production* (Beyer, 1992). Even the modern classic on the spoken language, Tournadre's *Manual of Standard Tibetan* (2003), is descriptive and presented in English (or French).

[27] Macaro (2006): http://goo.gl/egx3ic. "Our results support previous findings that explicit instruction [does not lead to]... gains in accuracy in either translation or free composition." Also see: Hastings (2004): http://goo.gl/AOEhCG.. "The researchers suggest that regular encounters with the real language—in other words, comprehensible input—is the true source of grammatical competence."

[28] See: Ueda (1979: 78-103). Kenkyusha (1988: 45-55). See also: Ogawa (2011).

> Contemporary texts for the teaching of foreign languages at the college level often reflect Grammar-Translation principles. These texts are frequently the product of people trained in literature rather than in language teaching or applied linguistics. Consequently, though it may be true to say that the Grammar-Translation Method is still widely practiced, it has no advocates. It is a method for which there is no theory. There is no literature that offers a rationale or justification for it or that attempts to relate it to issues in linguistics, psychology, or educational theory (Richards [2001: 7]).

Furthermore, Grammar-Translation's method implies that *reading* can be developed in a vacuum, completely separate from *speaking*, *listening*, and *writing* skills. The problem here is that *reading* is actually directly linked to *listening*, in that readers recover the meaning of language phonologically; and *listening* has further links to *speaking*, in that comprehension and production are not distinct mental processes, but shared neural networks.[29] Real interpretive skills are built wholistically, on all of these abilities together, and a modern pedagogy for the Tibetan language needs to take this reality seriously; restricting language study to grammar and vocabulary, without developing any productive or phonological skills, not to mention automaticity,[30] does not give students the opportunity to develop true reading comprehension, nor the ability to understand the language with accuracy or fluency.[31]

It's important to note, too, that the "classical" language of Tibetan doesn't date back to antiquity; and, like modern Icelandic speakers who can read 1,000 year-old Old Norse epics, or modern Arabic speakers who use MSA, educated native speakers of modern Tibetan can read, understand, and write in literary Tibetan.[32] It exists within the sociocultural context of a living oral tradition that works, in essence, by bridging formal and informal language in domains of *both* speech and text.[33] A pedagogy divorced from these factors is severely lacking, especially when one considers recent research that suggests language is a function of the social brain.[34] In other words, the creation of meaning depends, inextricably, on a shared mode of discourse, on a lived experience of language that is socioculturally embedded.[35] We may add to all these points that fluency in spoken Tibetan adds the potential for deep and nuanced collaboration with emic scholars.

[29] See: Treiman et al., (2003: 527-548): http://goo.gl/j7Eh9. Macdonald (2013): http://goo.gl/al4G2w. Pickering (2013): http://goo.gl/al4G2w. "Language production's impact on language comprehension is so pervasive that understanding production is essential to understanding comprehension."

[30] See: Lantolf (Ed.). (2000b). To paraphrase: Readers must make use of long-term memory to understand reference, relevance, and implication in order to then understand how sentences are integrated into a larger causal structure; readers must do this by analyzing events in terms of goals, actions, and reactions, and to do this, they must be able to take in vast amounts of information quite rapidly (automaticity). On discourse-level mental fatigue in grammar-translation methodology of Japan, see: Norris (1994: 25-38).

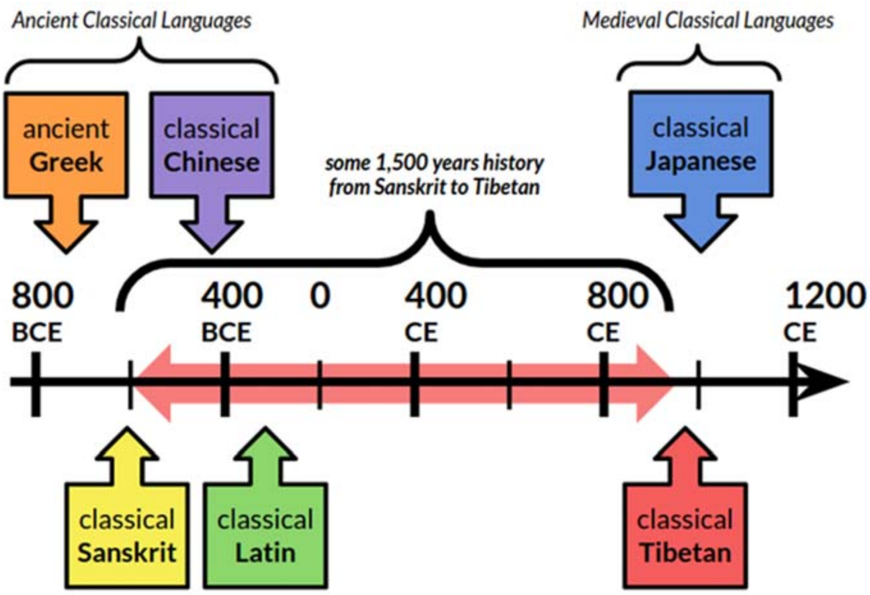
[31] Doughty (1998).

[32] Sanders, Ruth (2010: 209). Tournadre (2003: 26). Literary and spoken Tibetan "share the same basic grammar and are very similar lexically... with a knowledge of one it is possible to read the other without too much difficulty."

[33] Klein (1994: 281-314). Also see: Tournadre (2003: 27), where he notes that while literary Tibetan is not generally used for conversation, "some lamas or lay intellectuals use a form of expression which is virtually Literary Tibetan… there is therefore a real diglossia in their speech."

[34] rabe, p. 377, and p. 381 - "Cultural knowledge has been shown to influence comprehension." Ibid., p. 388. "Linguistic differences at syntactic and discourse levels are more likely to have an influence on reader comprehension." Readers of different languages pay attention to different types of words—languages encode information differently both syntactically and organizationally. See also research by Kuhl (2007): http://goo.gl/kFsOce & Thomson (2006): https://goo.gl/BU1F6s.

[35] Bruner (1990). Lantolf (2000).



The misleadingly labeled "classical" era of Tibetan literature is separated from the classical languages of antiquity by well over 1,000 years of history.

Finally, while the classicist's approach of Grammar-Translation *might* be applied with some success to the starkly inflected Indo-European languages, like Sanskrit, which have grammatical cases very clearly demarcating the relationships between words within a sentence, it seems fair to question whether or not it's reasonable to expect similar results for Tibetan, which is isolating, weakly inflected, and on the whole more ambiguous in nature.[36] All this to say that the most efficient pedagogical path toward Tibetan reading comprehension for L2 students may actually be first grounding students in the spoken language[37]—just as the field of teaching Arabic as a foreign language has discovered[38]—while building up to the full breadth of sophisticated textual discourse by gradually filling in gaps left by the diglossia based on sound linguistic research. The following section aims to show exactly how the process of applying corpus data to educational materials can inform curriculum development.

3.1 Curriculum development: A Tibetan alphabet book

As discussed, frequency data is essential for educationalists who desire an effective, research-based pedagogy. To demonstrate how, we may use letter frequency to inform the order of lessons for

[36] See: Tournadre (2010). Since Tibetan grammar is traditionally described using the Sanskrit case system, one can immediately see that, in comparison: nouns are not marked by gender or number; grammar is particle-based, not inflective; the first case has no grammatical marking; the *ladön* stretches over three Sanskrit case meanings; and the vocative case doesn't truly exist in Tibetan. Even though Tournadre divides literary grammar over 10 cases (or 2 "super cases"), he also asserts that they are "transcategorical" and "optional" in nature.

[37] See essays in: Gelder (1995).

[38] Especially see: Younes (2015). See again: Al-Mamari (2011). See also: Trentman (2011): http://goo.gl/74KVL5. See also again: Maamouri (1998). Finally, see: Sneed, Carol (2012): https://goo.gl/hOSJkD.

a Tibetan Alphabet Book just as we would use corpus data (vocabulary and syntactical frequency) to inform the larger scale of textbook creation (or the grading of level-specific reading material). Here, we begin by noting that traditional pedagogy for learning alphabetic principles focuses on introducing the alphabet a letter at a time in a traditional order—the "ABCs" for English speakers; "ཀ་ཁ་ག་ང་།" (*ka kha ga nga*) for Tibetan.[39] What educational research has shown, however, is that this form of education can actually be disadvantageous to the beginning reader.[40] Instead, students benefit most from explicit instruction that concretely makes the connection between sound (phoneme) and text (grapheme).[41] We're able to make these connections explicit by:

1. using an **appropriate print size** (and avoiding cursive);[42]
2. using **bold, color-coded text-to-sound connections**;
3. using **word separation;**[43] and
4. consulting **frequency analysis** and other linguistic research.[44]

Research also shows that beginning students learn more effectively when alphabetic principles are introduced according to what letters and groups of letters they will most frequently encounter[45] and begin reading quicker when introduced to words they will frequently read;[46] these principles remain true for readers of all levels.[47] By modeling a systematic set of introductory lessons on the Tibetan alphabet using these (and other) modern pedagogical principles, my aim was giving the student of the language the best possible start for later literary achievement.[48] Below, you may view the general color-coded schematic for the basic types of sounds found in Tibetan, organized along principles that unite classical emic descriptions of Tibetan phonology and modern linguistics:[49]

[39] Bowman (2004: 295–303).

[40] Piasta (2010: 8–38).

[41] Jones (2012).

[42] Print is easier than cursive: Graves (2010). Larger CPS is easier than smaller: Leggen (2011). Capital letters are easier to read: Smythe (1971).

[43] See Tenzin Dickyi (2010): http://goo.gl/Z5C6PA. Testing separated and unseparated text amongst Tibetans, separated text was, without exception, more easily comprehended. Also see: Saenger (1997). Also note that, in Japanese, "Interword spacing facilitated both word identification and eye guidance when reading syllabic script," Saiano (2007): http://goo.gl/2Bj3To; and in Thai, "Results show that subjects were faster in reading and made less errors when spaces were added," Chananda Kohsam (1997): http://goo.gl/S8cK24 (note that there *is* some research suggesting these same conclusions may not apply to ideogramic languages such as Chinese, except for non-native learners: see, for example, Bassetti: http://goo.gl/whBr1r).

[44] Gibson (1975). Treiman (1994: 97–122).

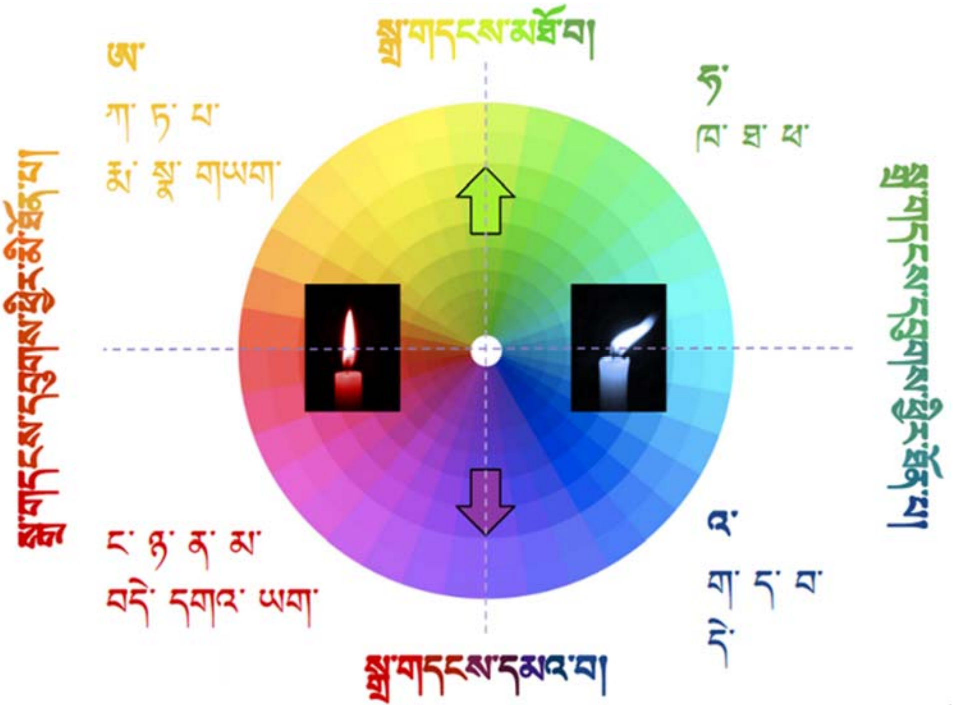
[45] Treiman (1994).

[46] Daud, Nuraihan Mat, et al. (2013).

[47] Murphy (1947). Swanson (1948).

[48] NELP (2008).

[49] The traditional ང་རོའི་རྩ་བ་གསུམ།, "the 3 roots of pronunciation." Refer to ཨ་ལག་ཤ་བསྟན་དར་ལྷ་རམས་པ། or བོ་གསལ་དགའ་སྟོན། by དམུ་དགེ་བསམ་གཏན།.

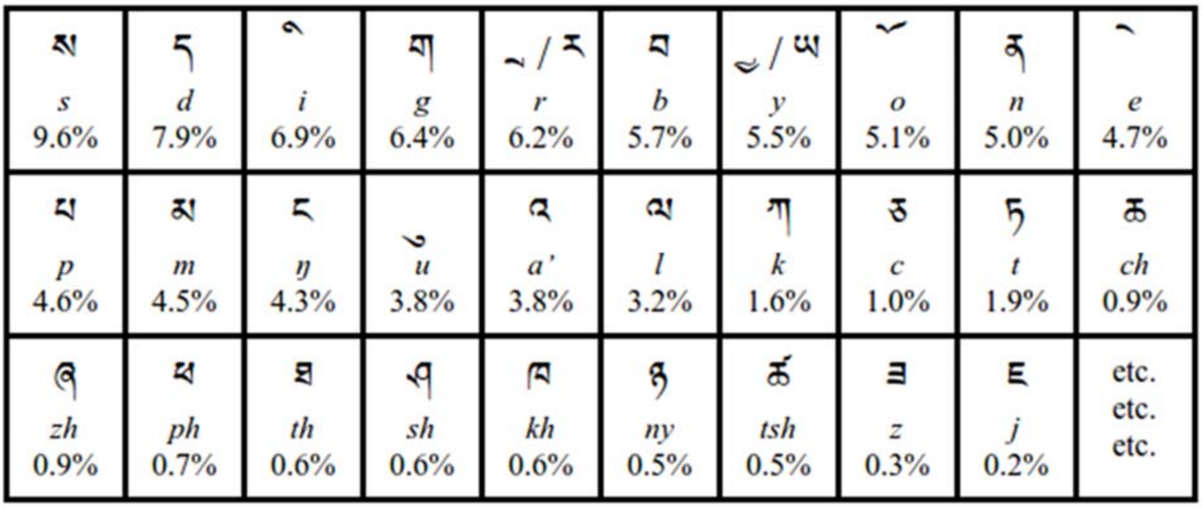


Above: The sounds of Tibetan are divided left-to-right by ***unaspirated*** *(warm colors to denote a candle continuing to burn) and* ***aspirated*** *(cool colors to symbolize the breath flickering the candle); similarly, sounds are divided top-to-bottom by* ***high and tight*** *(light colors) and* ***low and loose*** *(dark colors).*[50]

As for consulting corpus data for frequency information, we may use: Huidan Liu et al's modern text corpus and analysis (2014); Cai Zhi Jie et al's modern text corpus (2013); and my own analysis of traditional literature, which utilized a python-based letter and syllable counter on the digital input of the བཀའ་འགྱུར་ (*bka' 'gyur*) section of the Tibetan Buddhist Canon.[51] Upon analysis, these multiple corpora data points seemed consistent enough across both modern and classical literary Tibetan to be sufficient for understanding the basic frequency of letter distribution for an Alphabet Book; as aforementioned, later lessons and edits will require more data, namely frequency not only of letters, but of words; collocates; phrases; and more.

[50] Although developed independently, this scheme is similar to one developed by Tenzin Norbu Nangsal, and I have borrowed his headings for the quadrents in the graphic above. For Tenzin-la's system, see: https://www.youtube.com/watch?v=qZCcsICgC3s

[51] Programmed by Esukhia's Hélios Hieldt (http://esukhia.org/). Surprisingly (or not), this data showed very little variation from the modern text analyses cited above. Get the raw data here: https://goo.gl/Nl27VJ.



ས *s* 9.6%	ད *d* 7.9%	ི *i* 6.9%	ག *g* 6.4%	ྲ / ར *r* 6.2%	བ *b* 5.7%	ྱ / ཡ *y* 5.5%	ོ *o* 5.1%	ན *n* 5.0%	ེ *e* 4.7%
པ *p* 4.6%	མ *m* 4.5%	ང *ŋ* 4.3%	ུ *u* 3.8%	འ *a'* 3.8%	ལ *l* 3.2%	ཀ *k* 1.6%	ཅ *c* 1.0%	ཏ *t* 1.9%	ཆ *ch* 0.9%
ཞ *zh* 0.9%	ཕ *ph* 0.7%	ཐ *th* 0.6%	ཤ *sh* 0.6%	ཁ *kh* 0.6%	ཉ *ny* 0.5%	ཚ *tsh* 0.5%	ཟ *z* 0.3%	ཇ *j* 0.2%	etc. etc. etc.

Sample of Tibetan Letters, by order of frequency

When developing the introductory order of the alphabetic symbols, I also had three important questions in mind:

1. Is it **graphemically frequent** (does it occur frequently in text and speech)?
2. Is it **phonemically consistent** (is it always or almost always pronounced the same way)?
3. Is it **useful** for terms found in everyday speech? (can it be concretely connected to a **communicative purpose**)?

In other words, spellings and sounds that are *frequent*, *consistent*, and *useful* had priority; these were introduced first. The further the sounds and spellings were to these foundational principles, the later they came in the instructional cycle. By introducing small groups of letters at a time; showing how the sounds combine to form short, simple words (that are both useful and actually exist); avoiding similar sounds and shapes within one lesson; and practicing with both spoken and written activities to provide non-repetitive repetitions, I ended up with a new order of alphabetic introduction which looks like this:[52]

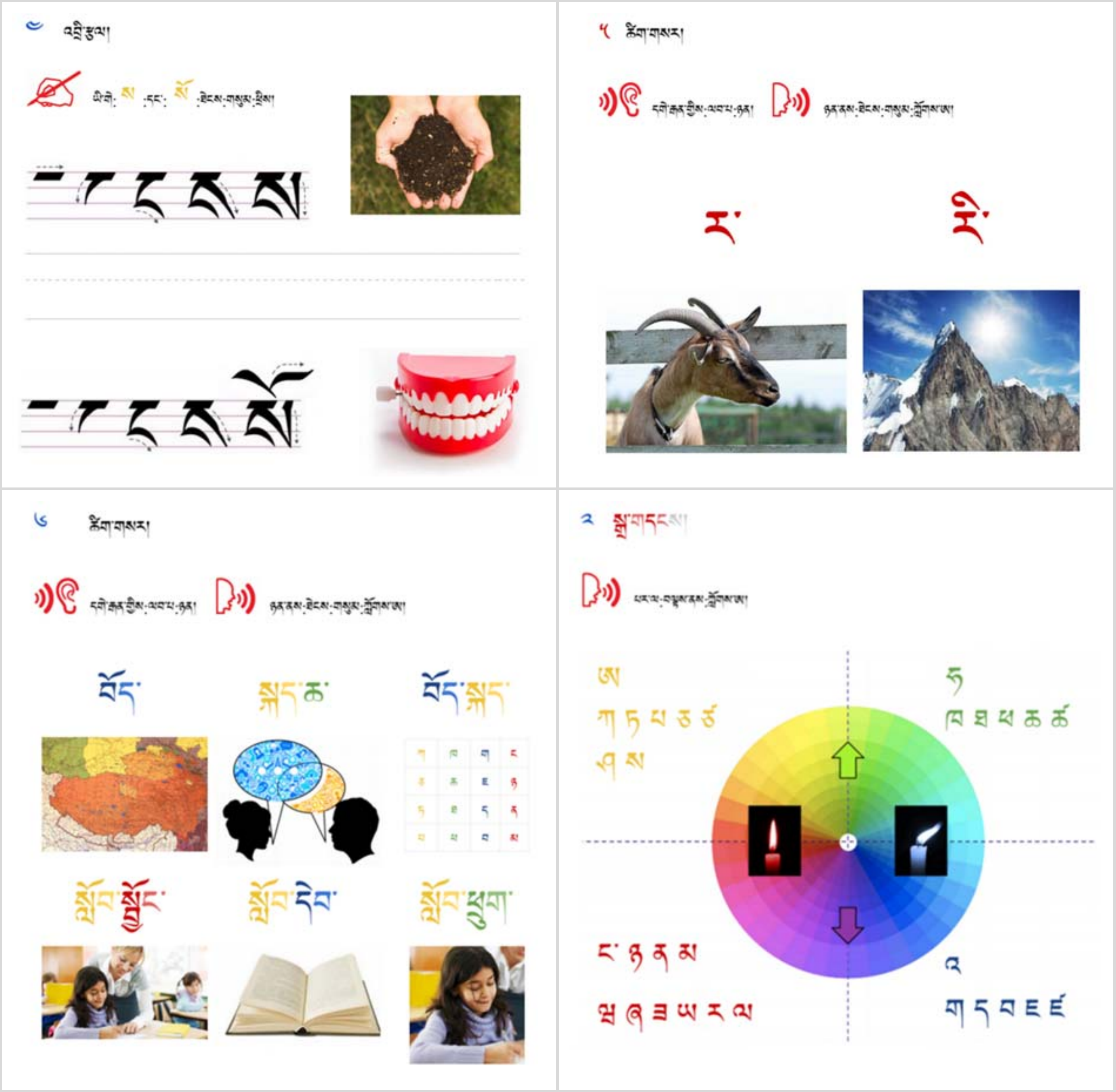
Lesson:	Letters	Vowels	Lesson Objectives
Lesson 1	ས ར མ བ	སོ རི མེ བུ	Common sounds and letters; basics of syllables; implicit "a" and vowel markers; basic punctuation, *tsheg* and *shad*; basic one-letter/one-vowel vocabulary
Lesson 2	ག ད འ		Basic words based in the alphabetic principle; sounded suffixes; silent prefixes

[52] See, for example, "How to Teach the Alphabet (ESL):" http://goo.gl/9Whq7g.

			and silent post-suffix "sa"; prefix effects on pronunciation
Lesson 3	ག ད བ □ ང		Headletter effects on pronunciation; basic vocab and sentence structure; "magic e" suffixes "sa" and "da"
Lesson 4	ཁ པ ཐ ཤ ན ལ		Aspirated letters; practice recognizing the 4 main sounds; subscribed "ra" and "ya"; basic spelling; pronouns and more simple sentences; simple conversation
Lesson 5	ཅ ཆ ཇ ཞ ཏ ཨ	/ö/ and /ɪ/	Practicing unaspirated/aspirated/voiced sounds ("a" "ha" "'a"); the un-alphabetic "བྱ" series; practicing vowel sounds /ö/ and /ɪ/ (the effect of "sa" and "da" on vowel sounds); basic verbs
Lesson 6	ཙ ཚ ཛ ཉ ཟ ཡ		Adding the alveolar affricates; exceptions to the alphabetic principle "ཟླ" and "སྦྱ"; traditional alphabet and spelling; basic word formation; review

Although I present the lesson plans here in English, the lessons themselves use *no* first language reference points, allowing a student to learn Tibetan in a purely Tibetan context from the very first day onward. The goal for language learning is to promote **exposure** to the language (ideally native speech) and encourage **production** in an immersion environment in order to elicit native-like understanding; a Tibetan-only context is key in maximizing these research-based principles.[53] By avoiding the ambiguous and confusing, and providing immediate and explicit connections to meaning and sound, the hope is that students who study beginning with the Alphabet Book, སོ་རི་མེ་བུ། (*so ri me bu*), will be provided with a concrete stepping-stone to further literary achievement in the Tibetan language.

[53] Morgan-Short (2012). Doughty (1998).



Above: Screenshots from the beta-release of "so ri me bu"; students learn in a Tibetan-only, color-coded format that makes explicit sound-to-meaning connections, while practicing all four languages skills of speaking, listening, reading, and writing.

4 Concluding remarks

In brief, work in corpus linguistics is vital to the future of Tibetan language education and literacy. Even the most rudimentary corpus information can provide invaluable frequency data to guide pedagogists, educationalists, authors, journalists, and children's book writers in their quest for evermore effective reading materials for both L1 and L2 audiences. Beyond tools like text analysis software built from readability formulas or frequency lists for curriculum development and level assessment, future comparative work will help uncover the distinct nature and relationship between spoken and literary Tibetan. This will be key in bridging the gaps left by diglossia in order to formulate working strategies to bring students from low-level texts with a high overlap in spoken language to true reading comprehension for even the most sophisticated and elite levels of traditional Tibetan literature—while in the meantime deepening the everyday connection between people and texts through literacy and language education.

References

Allen, John. 2003. *The BBC News styleguide*. BBC Training and Development. Retrieved from http://www2.media.uoa.gr/lectures/linguistic_archives/academic_papers0506/notes/stylesheets_3.pdf

Al-Mamari, Hilal. 2011. "Arabic diglossia and Arabic as a foreign language." *Capstone Collection Paper* 2437. Brattleboro: SIT Graduate Institute.

Bassetti, Benedetta. "Interword spacing in Chinese reading." London: University of London.

Beaglehole, Velma J. 2010. "The full stop effect: using readability statistics with young writers." *Journal of Literacy and Technology* 11.4: 55-83.

Beyer, Stephan V. 1992. *The Classical Tibetan language*. Albany: SUNY Press.

Bowman, M.; and Treiman, R. 2004. "Stepping stones to reading." *Theory into Practice* 43: 295–303.

Bruner, Jerome S. 1990. *Acts of meaning*. Cambridge: Harvard University Press.

Cai Zhi Jie, et al. 2013. "Attribute analysis in basic components of Tibetan word." *International Journal of Intelligent Engineering and Systems* 6.2: 1-7.

Callander, Nichola. 2011. *Communication, language, and literacy*. New York: Bloomsbury Academic.

Coleman, Meri; and Liau, T. L. 1975. "A computer readability formula designed for machine scoring." *Journal of Applied Psychology* 60.2: 283-284.

Daud, Nuraihan Mat et al. 2013. "A corpus-based readability formula for estimate of Arabic texts reading difficulty." *World Applied Sciences Journal* 21: 168-173.

Doughty, Catherine; Williams, Jessica, eds. 1998. *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press.

DuBay, W. H. 2006. "Smart language: Readers, Readability, and the Grading of Text." Costa Mesa: Impact Information.

Flesch, Rudolph. *How to write plain English*. Christchurch: University of Canterbury. Retrieved from http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

Ferguson, C.A. 1959. "Diglossia." *Word* 15: 325–340.

Gelder, Beatrice 1995. *Speech and reading: A comparative approach*. Erlbaum (UK): Psychology Press.

Gibson, E. J.; and Levin, H. 1975. *The psychology of reading*. Cambridge, MA: MIT Press.

Graves, Bill. 2010. "Most college students print as cursive writing starts to disappear." The Oregonian. Retrieved from http://www.oregonlive.com/education/index.ssf/2010/10/most_college_students_print_as.html

Hamilton, C.; and Shinn, Mark R. 2003. "Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills." *School Psychology Review* 32: 228-240.

Harmer, Jeremy. 2007. *How to teach English*. Edinburgh Gate, England: Pearson Education Ltd.

Hastings, Ashley; and Murphy, Brenda. 2004. "Implicit standards for explicit grammar teaching." Retrieved from http://www.focalskills.info/articles/implicit.pdf

Hillocks, G., Jr.; and Smith, M. W. 1991. "Grammar and usage." In Flood, J.; Jensen, J. M.; Lapp, D.; and Squire, J. R. Eds.), *Handbook of research on teaching the English language arts,* 591-603. New York: Macmillan.

Huckin, Thomas. 1997. *Second language vocabulary acquisition*. Cambridge: Cambridge University Press.

Liu, Huidan, et al. 2014. "Zipf's Law and statistical data on Modern Tibetan." Chinese Academy of. Sciences, Beijing,. China. pp. 323-333. Retrieved from http://www.aclweb.org/anthology/C14-1032

Jones, C. D.; and Reutzel, D. R. 2012. "Enhanced alphabet knowledge instruction: Exploring a change of frequency, focus, and distributed cycles of review." *Reading Psychology* 33.5.

Klare, G. R.; and Buck 1954. *Know your reader: The scientific approach to readability*. New York: Heritage House.

Klein, Anne Carolyn. 1994. "Oral Genres and the Art of Reading in Tibet." *Oral Tradition* 9.2: 281-314.

Kohsam, Chananda, et al. 1997. "Adding Spaces to Thai and English: Effects on reading." Proceedings of the 19th Annual Meeting of the Cognitive Science Society, p. 388-393. Hillsdale, NJ: Erlbaum.

Kuhl, Patricia K. 2007. "Is speech learning 'gated' by the social brain?" *Developmental Science* 10.1: 110–120.

Lantolf, J. P. (Ed.). 2000. *Socio-cultural theory and second language acquisition*. Oxford: Oxford University Press.

Legge, Gordon E.; and Bigelow, Charles A. 2011. "Does print size matter for reading? A review of findings from vision science and typography." *Journal of vision* 11.5: 1-22.

Lostutter, M. 1947. "Some critical factors in newspaper readability." *Journalism quarterly* 24: 307–314.

Maamouri, Mohamed. 1998. *Arabic diglossia and its impact on the quality of education in the Arab region.* International Literacy Institute: University of Pennsylvania.

Macaro, Ernesto 2006. "Does intensive explicit grammar instruction make all the difference?" *Language Teaching Research* 10.3: 297–327.

Marsden, Richard. 2011. "The Bible in English in the Middle Ages". *The practice of the Bible in the Middle Ages: Production, reception and performance in Western Christianity.* New York: Columbia University Press: 272-295.

McEnery, Tony. "What corpora can offer in language teaching and learning." Lancaster: Lancaster University. Retrieved from http://lancs.ac.uk/~xiaoz/papers/Corpora%20and%20language%20teachingv7.rtf

McWhorter, John. 2013. "Is texting killing the English language?" TIME. Retrieved from http://ideas.time.com/2013/04/25/is-texting-killing-the-english-language/

Morgan-Short, Karen, et al. 2012. "Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns." *MIT Journal of Cognitive Neuroscience* 24.4: 933-947.

Murphy, D. 1947. "How plain talk increases readership 45% to 60%." *Printer's Ink* 220: 35–37.

National Early Literacy Panel (NELP). 2008. *Developing early literacy: Report of the national early literacy panel*. Washington, DC: National Institute for Literacy.

Nikolajeva, Maria. 2014. *Reading for learning: Cognitive approaches to children's literature.* Philadelphia: John Benjamins Publishing Co.

Norris, Robert W. 1994. "Getting students more personally involved: An alternative to the yakudoku- and lecture-dominated methods of teaching literature and reading." *Fukuoka Women's Junior College Studies* 48: 25-38.

Piasta, S. B.; and Wagner, R. K. 2010. "Developing early literacy skills: A meta-analysis of alphabet learning and instruction." *Reading Research Quarterly* 45: 8–38.

Pressley, Michael. 2006. *Child and adolescent development for educators*. New York: Guilford Press.

Richards, Jack C.; Rodgers, Theodore S. 2001. *Approaches and methods in language teaching (2nd ed.)*. Cambridge: Cambridge University Press.

Rockwell, John. 1991. *A primer for Classical Literary Tibetan.* Retrieved from https://learntibetian.files.wordpress.com/2012/08/rockwell-primer-for-classical-literary-tibetan.pdf

Saenge, Paul. 1997. *Space between words: The origins of silent reading*. Stanford University Press: Stanford, CA.

Saiano, Miia, et al. 2007. "The role of interword spacing in reading Japanese." *Vision Research* 47.20: 2575–2584.

Sanders, Ruth. 2010. *German: Biography of a language*. Oxford: Oxford University Press.

Schmidt, Dirk. 2014. "Rethinking Classical Tibetan pedagogy."
Retrieved from
https://www.academia.edu/12915197/Rethinking_Classical_Tibetan_Pedagogy

Smythe, P. C.; Stennett, R. G.; Hardy, M.; and Wilson, H. R. 1971. "Developmental patterns in elemental skills: Knowledge of uppercase and lower-case letter names." *Journal of Reading Behavior* 3:3: 24-33.

Sneed, Carol. 2012. "Arabic diglossia and Arabic language teaching: Teaching and learning vernacular and standard varieties." *Growing participator approach*. Retrieved from https://growingparticipatorapproach.wordpress.com/teaching-msa-and-colloquial-arabic/

Swanson, C. E. 1948. "Readability and readership: A controlled experiment." *Journalism quarterly* 25: 339–343.

Taylor, Alex. 2003. "Language teaching methods: An overview." Retrieved from http://blog.tjtaylor.net/teaching-methods/

Tenzin Dickyi. 2010. "Breathing space: How word separation can save the Tibetan language." Phayul. Retrieved from http://www.phayul.com/news/article.aspx?id=26482

Thomson, Greg. 2006. "Introduction to the sociocultural dimension of language learning." Accessed: 16 October 2015.

Tournadre, Nicolas. 2010. "The Classical Tibetan cases and their transcategoriality: From sacred grammar to modern linguistics." *Himalayan Linguistics* 9.2: 87-125.

Tournadre, Nicholas. 2003. *Manual of Standard Tibetan*. New York: Snow Lion.

Treiman, R.; Weatherston, S.; and Berch, D. 1994. The role of letter names in children's learning of phoneme-grapheme relations. *Applied Psycholinguistics* 15: 97–122.

Trentman, Emma. 2011. "L2 Arabic dialect comprehension: Empirical evidence for the transfer of familiar dialect knowledge to unfamiliar dialects." *L2 Journal* 3.1: 22-49.

Tunberg, Terence and Minkova, Milena 2012. "Active Latin: Speaking, Writing, Hearing the Language." *New England Classical Journal* 39.2: 113-28.

Waring, Rob; and Takaki, Misako. 2003. "At what rate do learners learn and retain new vocabulary from reading a graded reader?" *Reading in a Foreign Language* 15.2: 130-163.

Wan-a-rom, Udor 2008. "Comparing the vocabulary of different graded-reading schemes," *Reading in a Foreign Language* 20.1: 43–69.

Younes, Munther 2015. *The integrated approach to Arabic instruction*. London and New York: Rutledge. See again: Al-Mamari (2011).

Dirk Schmidt
thedirk@gmail.com