NLP for Readability, Graded Literature, & Materials Development in Tibetan*

Dirk Schmidt
(University of Wisconsin-Madison)

When it comes to learning-to-read, Tibetan is notoriously difficult. This is reflected in low literacy rates, low levels of reading comprehension, and struggles by early readers of all kinds—children, native speakers, heritage speakers, and second-language learners alike. The most recent statistics for the Tibetan Autonomous Region (TAR), for example, show rates of illiteracy from 21% to 34% (Reddy & Bhole 2023, Textor 2022). Beyer (1992) describes literacy in Tibetan as an ”elite achievement”. And earlier investigations into reading comprehension and vocabulary levels in diaspora contexts have found comprehension difficulties for Tibetan literature (Schmidt 2022a). At the core of this issue is diglossia, the gap between how Tibetan is spoken and how it is written (Ferguson 1959). Briefly put, the closer a writing system hews to a speech community's own natural language—the variety they use for everyday communication—the easier it is for speakers to learn to read. Correspondingly, the further apart speech and writing are, the more difficult literacy is (Koda, Zehler, Perfetti & Dunlap 2008).

In the case of Tibetan, written norms date to the 7th–11th centuries, and have changed little over the last 1,000 years (Tournadre & Gsang-bdag-rdo-rje 2003). This means that readers and writers must take effort to process their natural languages while decoding or encoding Tibetan text—mentally adding or subtracting letters that are no longer pronounced, replacing speech words for written corollaries, and making mental grammatical or syntactic substitutions or other changes. Tibetan text—even text written for early readers—is thus rarely highly readable. For example, following Nation's method for analysis (Hu & Nation 2000), a sample of 26 published children's stories was found to have an average readability of only 65%.[1] This is significantly lower

* Dirk Schmidt, “NLP for Readability, Graded Literature, & Materials Development in Tibetan”, *Revue d'Etudes Tibétaines*, no. 72, Juillet 2024, pp. 70-85.

With thanks to Esukhia.

[1] For this, I wrote Python code to import digital text from children's stories and calculate a readability percentage. To account for automation and segmentation errors, stories above 75% (rather than 98%) were deemed reasonably ‘readable‘. However, 9/10 stories still fell below this benchmark (Schmidt 2022b).

than the recommended vocabulary coverage of 98% for independent reading (ibid.).

This paper aims to show why this is the case, and what can be done about it. It builds on my previous work (Schmidt 2020), but is more expansive in scope, providing important updates, a thorough theoretical backing, and more technical details about the role of *Natural Language Processing* (NLP) to the work of readability. Specifically, I will cover the learning-to-read process step-by-step, and how this impacts early readers of Tibetan. Then, I will propose a new data collection technique for writing beginning reading material, the creation of story-specific 'mini speech corpora'. Finally, I will focus on how to apply these ideas to the Tibetan context using word segmentation in both *Dakje* (Esukhia 2022b) and *Botok* (OpenPecha 2023) for editing and level identification. My aim is to provide readers who wish to write, edit, or analyze early reading materials with the practical information, tools, and resources they need to do so, in the hopes that it benefits and supports readers of the Tibetan languages.

1. Introducing Applications for NLP

While technical and theoretical NLP work for Tibetan began in the 1990s (Hill & Jiang 2016), some of the more practical, everyday applications for the field are just now becoming widely available, or are now in their nascent stages. To provide a brief overview of NLP tools that provide practical applications for everyday users, large tech companies and small initiatives have both played important roles. Google Cloud Vision now provides Tibetan *Optical Character Recognition* (OCR) (Google 2023), building on early progress made by Namsel (Rowinski 2016). Microsoft recently released *Machine Translation* (MT) for Tibetan (Lekhden 2021). Tools like speech recognition (Ruan, Gan, Liu & Guo 2017) and spell-checking (Roux 2017) have also seen progress. These developments have followed in the footsteps of progress made in majority languages, like English.

Many of them are also dependent on progress in fundamental NLP areas like word segmentation and POS tagging (Hill & Jiang 2016). Dakje and Botok, the Python segmenter it relies on for word spacing and recognition, follow this trend of modeling itself on progress made in the larger languages. Specifically, Dakje uses Botok's word segmentation to build on ideas in vocabulary analysis, grading (or leveling), and readability scores found in tools built for grading and editing text (Chall 1948). Writers who write in English, for example, may use an editor like Hemingway (Long 2023) to analyze, simplify, and improve the readability of their text. Similarly, Dakje provides a user-friendly interface for readability editing in Tibetan. For users with basic Python coding skills, Botok provides the user with more advanced options for segmenting text for readability analysis. After briefly outlining the background to issues in readability, this article will present how word segmentation in Dakje and Botok have been used in the context of editing and analyzing Tibetan text for early reading materials for Esukhia,[2] a non-profit organization I work with that creates resources for Tibetan language learning.

2. *The Issue: Diglossia*

To understand how word segmentation is key to improving readability, it is important to first discuss the big-picture context of learning to read in general. When we take a closer look at how readers attain literacy, the ways in which diglossia creates obstacles to literacy become clear (Hudson 1992, Harbi 2022). Correspondingly, the ways in which word segmentation in Dakje or Botok helps writers clear these obstacles should also become apparent. For the purposes of this article, I will divide the road to literacy into four steps (below). This road map is greatly simplified. Learning to read is a complex process, and many of these 'steps' occur in parallel, and inform one another. Below, each of these steps will be introduced and expanded, followed by a discussion of how word segmentation works to improve the learning-to-read process:

2.1 Developing speech skills & reading habits

2.2 Connecting sounds to symbols (& symbols to sounds)

2.3 Reading level-appropriate texts extensively

2.4 Vocabulary growth & learning from reading

2.1 *Developing speech skills & reading habits*

The first step on the road to literacy is developing speech skills and good reading habits. Children's (or a second-language learner's) exposure to oral language leads to the acquisition of speech skills. Meanwhile, being read to—out loud—creates motivation for reading, leading to good reading habits (Brock & Rankin 2008). Pressley & McCormick (2007) and others call this level ”emergent literacy skills”, while Callander & Nahmad-Williams (2011) draw important links between factors like early communication, rhythm, companionship, and social skills in early language development. The diglossic gap between speech and writing, however, make naturally-obtained speech skills

[2] https://esukhia.net/

less useful in this learning-to-read process. Known words appear less frequently, and books that contain unknown words won't be understood, which impacts motivation for further reading. Sevinç & Backus (2019) give more details on this kind of language anxiety in a specific context, which is also discussed more below.

2.2 *Connecting sounds to symbols*

Next, a beginning reader must internalize the alphabetic principle, or what is called 'phonemic awareness'. These are the connections speakers make between the sounds from their natural language and the symbols found on the page (Brock & Rankin 2008: p.203). By sounding things out, they decode written words into speech words in order to understand the text. With practice reading out loud, oral comprehension gradually becomes reading comprehension. That is, understanding speech is what leads to understanding text. But when spellings are 'opaque'—that is, when there is not a one-to-one correspondence moving from symbols to sounds—'decoding' text becomes increasingly difficult, blocking this process. Research shows that children who learn to read in 'transparent' orthographies, for example, learn to read faster than those who learn 'opaque' ones (Koda et al. 2008). When sounding-things-out is difficult, reading is difficult; if the word that is decoded is an unknown word, it won't be understood.[3]

Modern Tibetan languages are, generally speaking, 'opaque', rather than 'transparent'. This is especially true of the Central Tibetan dialects most frequently spoken and studied in the diaspora and in the West more broadly. While the diaspora varieties are widely conflated with Central Lhasa Tibetan (Tournadre & Gsang-bdag-rdo-rje 2003), they have several unique features (Schmidt 2022a). Here, I prefer the term Zhichag Tibetan (*gzhis-chags skad*, "settlement language"), and examples from these varieties are the ones referenced in this article. They broadly share many of the pronunciation features of the Central Tibetan dialects that lead to 'opaque' spellings, such as consonant cluster reduction. It is reasonable to expect, then, that this would have an effect on speakers and learners of these varieties learning to read or write it. To give an example of orthographic depth in Tibetan:

(1) Some Tibetan words are 'transparent' (they have a 1:1 symbol:sound relationship): In *ku-shu*, ཀུཤུ་, "apple", for example, all the consonants and vowels are pronounced as written, **/ku-çu/**.

[3] It's worth noting that sometimes, even a 'known' word won't be comprehended (Hu & Nation 2000).

(2) Many more words, however, are ‘opaque‘—they contain consonant clusters that are no longer pronounced; are pronounced differently than they are written; or are otherwise inconsistent in their grapheme-to-phoneme relationships: In *′bras*, འབྲས་ ”rice”, for example, the initial *′a* is silent; the cluster *-br-* has been palatalized; and the final *-s* changes the vowel, but is itself not pronounced, /ʈe/.

2.3 Reading level-appropriate texts extensively

With a foundation of speech skills and the ability to decode symbols into sounds, the third step in attaining literacy is reading level-appropriate texts extensively (Jacobs & Farrell 2012). In other words, getting considerable amounts of practice at reading. By reading simple, age- or level-appropriate texts extensively, readers build up their word-recognition skills (i.e., ‘automaticity‘). They improve in speed, fluency, and reading comprehension. In order for this to happen, reading material must be highly readable. Again, research suggests an ideal vocab coverage of 98% (Schmitt, Jiang & Grabe 2011, Hu & Nation 2000). Below, in Figure 1, "known" vocabulary refers to words in a speaker’s active vocabulary: Words they recognize, understand, and are able to use. A lack of easy, level-appropriate reading material that contains known words, however, puts a beginning reader at a disadvantage. Again, difficult texts can easily demotivate an early reader. A beginner in this situation is at risk of developing ‘language anxiety‘, a negative feedback loop that leads to less and less reading (Sevinç & Backus 2019). Competition from other, easier literatures, can also lead a reader to prefer using another language altogether for reading and writing (for example English or Chinese).

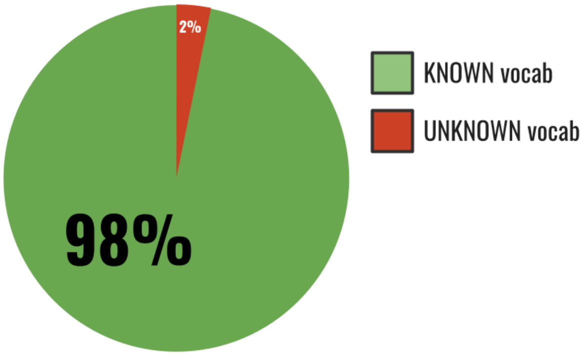


Fig. 1 – The target vocab coverage for a level-appropriate text is 98%

2.4 *Vocabulary growth & learning from reading*

It isn't until these foundational skills have been obtained that a reader can begin using literacy to learn from reading (Nikolajeva 2014, Wolf 2010). In step four, speech vocabulary grows as the beginning reader reads more and more. Understanding and enjoying the texts that they read leads to more reading, and more reading leads to better reading skills. In contrast, reading less and less—because it is too hard, too anxiety-inducing, or too impractical—leads to not being good at reading. Not being good at reading then leads, again, to less reading. Figure 2 provides a graphic view of this feedback cycle, where not comprehending a text causes frustration; frustration leads to decreased motivation; and less practice reading leads to less comprehension (which, in turn, leads to further frustration). The opaque spellings, high level of unknown vocabulary, and lack of level-appropriate reading materials of diglossia all contribute to this ”vicious cycle” of language anxiety (Sevinç & Backus 2019).

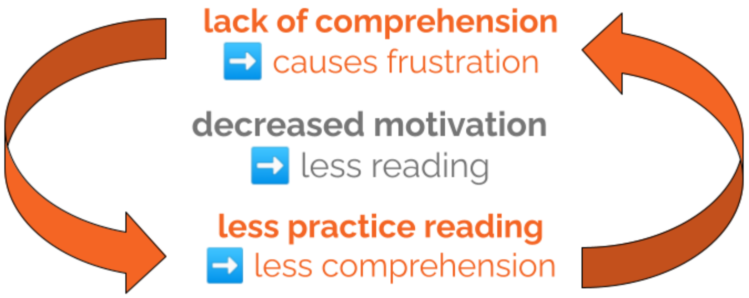


Fig. 2 – The negative feedback loop of 'language anxiety'

3. *The Solution: Readability*

Breaking this cycle requires easy-to-read texts that increase reading, motivation, and literacy. In the case of Tibetan, addressing issues like opaque spellings and shifts in grammar that have arisen from diglossia would require comprehensive language reforms, something that would take widespread social support and political will. Given the sociopolitical marginalization these languages face, this seems unlikely in the foreseeable future. However, vocabulary choice is something that any author, writer, or material developer can easily address. By choosing known words that occur naturally in speech over unknown ones, the readability of any given text can be greatly increased. And Natural Language Processing (NLP) using word-segmentation tools makes this not only possible, but easy and efficient. The following sections address how Dakje and Botok supports writing easy-to-read text by discussing each step in the development and implementation process of these tools:

3.1 Collecting natural speech (*putting the ’NL’ in ’NLP’*)

4.1 Processing the data (*adding the ’P’ to the mix*)

4.2 Writing level-appropriate texts (*using applied NLP to help writers*)

3.1 Collecting natural speech

For our purposes, then, it’s important to define natural language strictly in both time and space. That’s because the NLP that is useful in the context of readability is dependent on the speech community it will be used to benefit, or on the target language the learner is hoping to acquire. For Tibetan, this necessitates addressing the diglossic gap between speech and writing, as discussed above. But also requires recognizing that not all Tibetan speech communities use the same words and grammar in their natural speech—natural language is unplanned, naturally occurring, and constantly changing. Of the fifty or more Tibetan languages that exist, as defined by their mutual comprehensibility (Tournadre 2014), each have their own unique pronunciations, vocabularies, and grammars. To put it another way, a frequency list based on the words Zhichag Tibetan speakers use will not be the same as a list based on the words, say, Amdo Tibetan speakers use. Even if the target literature of Standard Literary Tibetan is the same, the early stepping stones may be different for different speakers and learners of different varieties. We need to know what words speakers know. This requires collecting natural speech data.

Methodologically, ‘collecting natural speech‘ means recording it using a voice recorder; transcribing it as it was spoken (that is, non-prescriptively, making no edits or corrections); and organizing the data for ease of analysis. For large corpus projects, the more data, the better. For smaller projects, however, we may use data that targets a specific demographic; a particular age group; or even an individual story. Creating these kinds of ‘mini corpora‘ for purposes of analysis and readability is one way to apply speech corpus creation to language learning and literacy (Beeching 2014, O’Keeffe, McCarthy & Carter 2007). One such example is the story ”The Race”, an open source children’s story from Pratham’s Storyweaver website (Figure 3).[4]

In our recent work creating mini speech corpora, we began by telling the story, orally, to children. Using the images (but no text), we then allowed

[4] https://storyweaver.org.in/

Fig. 3 – An example story, "The Race", is used to illustrate the feedback speech corpora can provide.

the children to re-tell the story back to us, in their own words, recording the result. In this way, we are able to limit the amount of data we have to collect, while ensuring we have story-specific vocabulary to work from. Afterwards, the recordings were transcribed, resulting in a mini speech corpus. This corpus contains the speech versions from several children of the same story (Esukhia 2022a). If the speech corpus researcher is also the writer of the Tibetan version, very little post-editing is needed. The result is a graded story, told in words that the children used (and thus, words that we can be sure they know). However, to make speech corpora more widely useful, further steps are useful for improved readability. The next section will explore how to apply this data using word segmentation in Dakje and Botok.

4. *The Path: NLP Tools*

So far, this article has introduced the problem: Tibetan texts have low readability due to diglossia. Texts contain opaque spellings, hard words, and literary grammar that do not occur in natural speech. It has also offered a solution: Readable texts for early readers that have a high percentage of known words from natural speech. The goal of using the NLP tools Dakje and Botok is identifying words that might be hard, or giving an overall sense of the grade level of a text based on its vocabulary. We do this by processing a text input, splitting it into words using NLP tools, and comparing those words to frequency lists made from speech data. Dakje gives the words a color based on how often they appear in natural speech. Similarly, Botok may be used within Python directly in much the same way, by segmenting text and comparing it to the word lists from natural speech. In this section, I will first discuss the details of this process; then, how it is applied in a real-world context in order to analyze, grade, or write children's stories or textbook materials.

4.1 Processing the data

The idea behind frequency lists is that the more often a word is used, the easier it is, and the more people are likely to know it. Hard words, in contrast, are rarer. They are used less often, and fewer readers are less likely to know them. Frequency has been used like this since the early days of graded reading to give researchers an idea of vocabulary difficulty (Chall 1948, DuBay 2007). For Tibetan, however, 'the word' is not an obvious unit: while the inter-syllabic Tibetan punctuation mark, or *tsheg*, indicates syllable boundaries, there is no punctuation that shows word boundaries. Ideally, we want to outsource the tasks of identifying, counting, and sorting Tibetan words to the machine. The result is word lists, ranked by frequency, that we can then split and sort into level lists.

In other words, here, a word's 'level' is defined by its frequency. While it is generally recognized that some words are easier and others harder, there is no universal, agreed-upon standard for precisely defining vocabulary levels or their lengths. For second-language learning, however, J & Alexiou (2009) provides basic guidance for length (or size) based on the Common European Framework of Reference for languages (CEFR). There, for example, 1,500 words is the suggested Beginner Level, or Level A1 (ibid). Combining this with the ideas from Chall (1948), among others, I have split and sorted words based on the principle that frequent words are easier and infrequent words are more difficult. This results in a set of lists that are used as reference points for vocabulary difficulty by level. In addition to this general data, we also have the story-specific vocabulary lists taken from the mini corpora collected for the stories.

4.2 Writing level-appropriate texts

As discussed in Section 3.1, writing or translating a beginning text directly (that is, without feedback on readability) will be successfully level appropriate if and only if the writer researches children's speech themselves. As shown in Figure 1, any percentage of unknown vocabulary beyond 2% is burdensome. For reference, this would be 9–10 unknown words every page in an article like this one. It's easy to see how unrecognizable words can lead to not reading when they make up even more than that. Imagine not knowing 20+ (5%), 40+ (10%), 60+ (15%) or even more of the words on each of these pages! In a ten-page article, you'd encounter hundreds of words you didn't know. In contrast, while it may seem counter intuitive, easy-to-read texts lead to more and better reading. This is why level appropriate texts are so important for literary achievement.

Perhaps surprisingly, translation of a level-appropriate story written in English, for example, does not automatically yield a level-appropriate Tibetan version (Schmidt 2020). That is because vocabulary choice in Tibetan writing is heavily influenced by many factors, including both traditional literary standards, as discussed above, as well as the movement for a modern "Pure Tibetan" (Tib. *bod-skad gtsang-ma*) (Thurston 2018). The fear that modern loanwords are 'degrading' Tibetan has led to large dictionary projects that collect, define, create, and publish Chinese-English-Tibetan dictionaries for new, modern vocabulary (Blo-gros 2013). Yet, while Tibetan children do use loanwords, the rate—even in the diaspora, amongst the youngest generations of speakers—does not seem to be particularly high. For example, if we analyze transcripts of diaspora children telling stories (Esukhia 2022a), we find that modern loanwords make up less than 1% of the total words spoken. This is represented in Figure 4. While a desire to preserve and promote Tibetan language is commendable, the impact of each additional unknown word can add up, overly burdening a beginning reader. The vocabulary choices from the many versions of "The Race" found on the StoryWeaver website exemplify this issue. Each of these vehicles has a specific neologism in Pure Tibetan, and this is reflected in the translated versions; the children, however, naturally used different vocabulary when speaking during corpus collection, suggesting they may not actually know or use these terms (see Table 1).

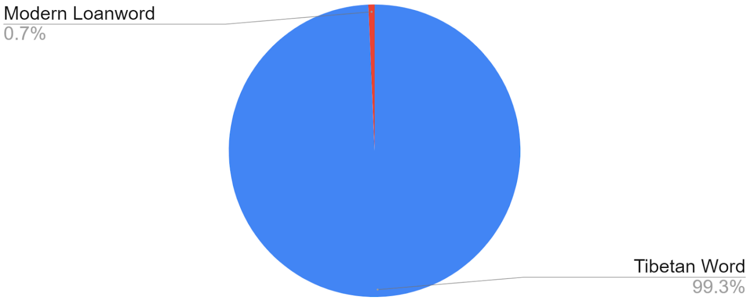


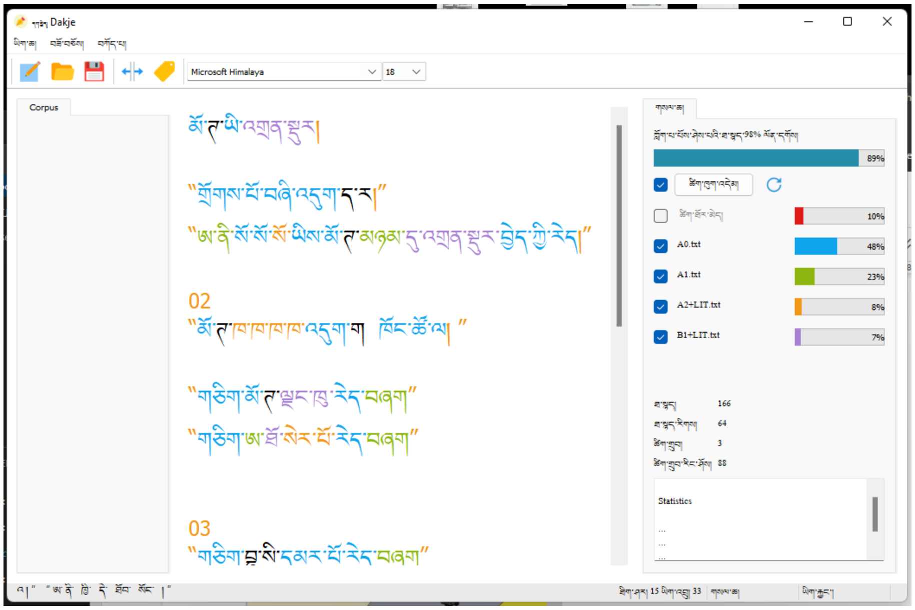
Fig. 4 – Even in diaspora children's speech, modern loanwords make up less than 1% of the total words spoken.

English	Pure Tibetan	Speech word
bus	*spyi-spyod-rlang-’khor,* སྤྱི་སྤྱོད་རླང་འཁོར་	*bus,* བྷ་སི་
auto rickshaw	*’khor-gsum-snum-’khor,* འཁོར་གསུམ་སྣུམ་འཁོར་	*auto,* ཨ་ཊོ་
car	*rlangs-’khor,* རླངས་འཁོར་	*mo-Ta,* མོ་ཊ་

Table 1 – The influence of ‘Pure Tibetan‘ on vocabulary choice in children’s stories; an example from ”The Race”. On the left, Pure Tibetan terms found in the published stories; on the right, natural speech loanwords found in the mini speech corpus.

4.3 Using Segmentation

As discussed above, segmentation is key to identifying non-level vocabulary, or unknown words beginning readers will find difficult. Whether done in Dakje (Esukhia 2022b) or Botok (OpenPecha 2023), the segmentation process relies on the same background processes. At its core, these tools implement a ”max match” algorithm for word recognition. Tibetan input text is compared to a large dictionary—in essence, a word list—and segmented based on matches to this list. In Dakje, the general word list is a dictionary of Standard Literary Tibetan. The benefit of using this software is that Dakje will automatically segment Tibetan text, and highlight vocabulary items by level. It will also calculate the distribution percentage across those lists (Figure 5, right panel), and display the total readability of the text (Figure 5, top bar). Users can then use this feedback to edit problematic words directly in the editor.



མོ་ཊ་ཡི་འགྲན་སྡུར།

"གྲོགས་པོ་བཞི་འདུག་ད་ར།"
"ཨ་ནི་སོ་སོ་སོ་ཡིས་མོ་ཊ་མཉམ་དུ་འགྲན་སྡུར་བྱེད་ཀྱི་རེད།"

02
"མོ་ཊ་ཁ་ཁ་ཁ་ཁ་འདུག་ག ཁོང་ཚོ་ལ། "

"གཅིག་མོ་ཊ་ལྗང་ཁུ་རེད་བཞག"
"གཅིག་ཨ་ཐོ་སེར་པོ་རེད་བཞག"

03
"གཅིག་ལྷ་སི་དམར་པོ་རེད་བཞག"

Fig. 5 – 'Dakje', an NLP-based word editing software for grading Tibetan texts, can provide actionable feedback for authors by highlighting words by level

The drawback, however, is that this the general dictionary lacks many words that are specific to speech in Modern Tibetan dialects. For those with coding ability, however, Botok allows the user to define a 'dialect pack' to improve word spacing. In the case of editing specifically for Zhichag Tibetan, for example, I prepared a word list of speech words from two Esukhia speech corpora: The Nanhai Corpus (Esukhia 2020) and The Children's Speech Corpus (Esukhia 2022a). I then loaded the speech word list into the dialect pack's "words" folder to call it when word spacing. I then segmented text by configuring Botok's Python module as below in Table 2. After word-spacing a text using this method, I then loaded the frequency lists as Python lists. This allowed me to grade texts directly by comparing them against the frequency lists. While this improved method still doesn't word-space perfectly, it is good enough for practical applications. For example, it allowed me to automatically assign levels to Esukhia's story database (Esukhia 2023). The database currently contains 42 stories, including five levels, L0–L4. These levels are roughly equivalent to the CEFR levels A0–B2, graded by word lists based on CEFR numbers (J & Alexiou 2009) and rates of unknown vocabulary (Nation & Hirsh 2020).[5] With a rise in the amount and quality of children's literature

[5] These ideas have also played a role in other materials development. See, for example, stories, games, and textbooks found on Esukhia's website.

in Tibetan, the hope is that these tools will reach a wider range of authors, writers, and material developers. With more, and more readable, content, beginning readers should have more opportunities that help them along the path to literacy.

```
def word_split(text):

  ”””takes a string of text and word spaces it based on the Zhichag dialect pack”””

    words=[]
    if __name__ == ”__main__”:
        config = Config(dialect_name=”zhichag”, base_path= Path.home())
        wt = WordTokenizer(config=config)
        tokens = get_tokens(wt, text)
        for token in tokens:
            words.append(token[’text’])

    return ’ ’.join(words)
```

Table 2 – A coding sample: Creating a user-defined 'dialect pack' for use in the Botok Python module.

5. *Concluding Remarks*

Using Dakje and Botok segmentation in the context of readability and applied linguistics is thus an important application for NLP in Tibetan. It can help authors, writers, and material developers ensure that they are providing their students, children readers, or language learners level-appropriate materials that help them develop good reading habits; connect symbols to sounds; and read extensively. Because diglossia manifests as unknown, difficult traditional or ”pure” vocabulary in beginning reading materials, NLP methods and tools like Dakje and Botok have an important role to play in breaking the cycle of language anxiety that comes hand-in-hand with attempts to attain literacy in diglossic languages. My hope is that the details provided in this article will support and encourage others to explore these tools to improve the readability of their own children’s stories and Tibetan language learning materials, too.

Bibliography

Beeching, Kate. 2014. Corpora in language teaching and learning. *Recherches en didactique des langues et des cultures* 11(1). doi:10.4000/rdlc.1672.

Beyer, Stephan V. 1992. *The classical tibetan language*. Albany: Suny Press.

Blo-gros, Tshul-khrims. 2013. *Rgya-bod-dbyin-gsum gsar-byung rgyun-bkol ris-'grel ming-mdzod[chinese-tibetan-english illustrated dictionary of new daily vocabulary]*. So-khron Mi-rigs Dpe-skrun-khang[Sichuan Nationalities Publishing House].

Brock, Avril & Carolynn Rankin. 2008. *Communication, language and literacy from birth to five*. Los Angeles: SAGE.

Callander, N. & L. Nahmad-Williams. 2011. *Communication, language and literacy* Supporting Development in the Early Years Foundation Stage. Bloomsbury Academic.

Chall, Dale E. 1948. A formula for predicting readability. *Educational Research Bulletin* 27. 11–20+28.

DuBay, William H. 2007. *Smart language: readers, readability, and the grading of text*. Costa Mesa, Calif.: Impact Information. OCLC: 164437606.

Esukhia. 2020. The nanhai corpus. https://github.com/Esukhia/Corpora/tree/master/Nanhai.

Esukhia. 2022a. Children's stories speech corpus. https://github.com/Esukhia/Corpora/tree/master/Childrens_Stories.

Esukhia. 2022b. Dakje. https://github.com/Esukhia/dakje-desktop.

Esukhia. 2023. Stories. https://esukhia.online/stories/.

Ferguson, Charles A. 1959. Diglossia. *WORD* 15(2). 325–340. doi:10.1080/00437956.1959.11659702.

Google. 2023. Ocr language support. https://cloud.google.com/vision/docs/languages.

Harbi, Mohammed. 2022. Arabic diglossia and its impact on the social communication and learning process of non-native Arabic learners: Students' perspective. *SSRN Electronic Journal* doi:10.2139/ssrn.4037655.

Hill, Nathan W. & Di Jiang. 2016. Tibetan natural language processing. *Himalayan Linguistics, Vol. 15(1)* .

Hu, Marcella & Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13.

Hudson, Alan. 1992. Diglossia: A bibliographic review. *Language in Society* 21(4). 611–674. http://www.jstor.org/stable/4168395.

J, Milton & T. Alexiou. 2009. *Vocabulary size and the common european framework of reference in languages* 194–211. Macmillan.

Jacobs, George M. & Thomas S. C. Farrell. 2012. *Teachers sourcebook for extensive reading*. Charlotte, N.C: Information Age Pub.

Koda, Keiko, Annette Marie Zehler, Charles A. Perfetti & Susan Dunlap. 2008. *Learning to read: General principles and writing system variation* 13–38. Routledge.

Lekhden, Tenzin. 2021. Microsoft’s translation app includes Tibetan language. https://www.phayul.com/2021/12/03/46489/.

Long, Adam Ben. 2023. Hemingway editor. https://hemingwayapp.com/.

Nation, Paul & David Hirsh. 2020. What vocabulary size is needed toread unsimplified texts for pleasure? doi:10.26686/wgtn.12560417.v1.

Nikolajeva, Maria. 2014. *Reading for learning: cognitive approaches to children’s literature* (Children’s literature, culture, and cognition v. 3). Amsterdam ; Philadelphia: John Benjamins Publishing Company.

O’Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom: language use and language teaching*. Cambridge ; New York: Cambridge University Press. OCLC: ocm76935901.

OpenPecha. 2023. Botok. https://github.com/OpenPecha/Botok.

Pressley, Michael & Christine B. McCormick. 2007. *Child and adolescent development for educators*. New York: Guilford Press. OCLC: ocm67383559.

Reddy, Rahul K. & Omkar Bhole. 2023. Analysing China’s Census Report .

Roux, Elie. 2017. Hunspell: Tibetan spellchecker. https://github.com/eroux/hunspell-bo.

Rowinski, Kurt, Zach;Keutzer. 2016. Namsel: An optical character recognition system for tibetan text. https://escholarship.org/uc/item/6d5781k5.

Ruan, Wenbin, Zhenye Gan, Bin Liu & Yinmei Guo. 2017. An improved tibetan lhasa speech recognition method based on deep neural network. *2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA)* 303–306.

Schmidt, Dirk. 2020. Grading tibetan children’s literature: A test case using the nlp readability tool “dakje”. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19(6). doi:10.1145/3392046. https://doi.org/10.1145/3392046.

Schmidt, Dirk. 2022a. *On the yak horns of a dilemma: Diverging standards in diaspora tibetan* 127–156. Amsterdam University Press. doi:10.1017/9789048552719.006.

Schmidt, Dirk. 2022b. Tibetan readability –a python project. https://github.com/thedirk/readability/blob/main/p00.ipynb.

Schmitt, Norbert, Xiangying Jiang & William Grabe. 2011. The percentage of words known in a text and reading comprehension. *The Modern Language Journal* 95(1). 26–43. http://www.jstor.org/stable/41262309.

Sevinç, Yeşim & Ad Backus. 2019. Anxiety, language use and linguistic competence in an immigrant context: a vicious circle? *International Journal of Bilingual Education and Bilingualism* 22(6). 706–724.

doi:10.1080/13670050.2017.1306021.

Textor, C. 2022. Illiteracy rate in China in 2021, by region. https://www.statista.com/statistics/278568/illiteracy-rate-in-china-by-region/.

Thurston, Timothy. 2018. The purist campaign as metadiscursive regime in china's tibet. *Inner Asia* 20(2). 199 – 218. doi:https://doi.org/10.1163/22105018-12340107.

Tournadre, Nicolas. 2014. *The Tibetic languages and their classification* 105–130. Berlin, Boston: De Gruyter Mouton. doi:10.1515/9783110310832.105.

Tournadre, Nicolas & Gsang-bdag-rdo-rje. 2003. *Manual of standard Tibetan: language and civilization: introduction to standad Tibetan (spoken and written) followed by an appendix on classical literary Tibetan*. Ithaca, N.Y: Snow Lion Publications.

Wolf, Maryanne. 2010. *Proust and the squid: the story and science of the reading brain*. New York: Harper Perennial 1st edn.