A Prosodic Lexicon of Lhasa Tibetan: An Experimental Study Based on Speech Synthesis

Authors: Lu Chen, Zu Yiqing, Liu Chenning, Zhang Xiao, Zu Yiqing

Date: 2025-08-07T00:00:00+00:00

Abstract

This study proposes a method for constructing a prosodic lexicon for Lhasa Tibetan based on a continuous speech database, applicable to low-resource, complex languages. The prosodic lexicon built from a small amount of high-quality data (3.77 hours, 2,526 sentences) significantly improves Lhasa Tibetan speech synthesis performance. The research reveals that linguistic components involved in tone sandhi in Lhasa Tibetan continuous speech are constrained by semantic, morphological, and syntactic factors, reflecting the hierarchical organization and chunking rules of the language cognitive system. The phonetic manifestations of tone sandhi components include three types: monosyllabic single-character tone, monosyllabic tone loss, and disyllabic tone sandhi. Which specific manifestation a syllable exhibits in connected speech is subject to three constraint conditions: the first type constrained by morphological rules, the second by syntactic rules, and the third by high-frequency grammatical constructions. Based on AI speech synthesis experiments, this study employs the first and third types of tone sandhi components and their constraint rules in continuous speech as the foundation for constructing the prosodic lexicon when building the language model, rather than conventional Tibetan dictionaries and word segmentation rules for information processing. Inspired by the "Usage-BasedTheory" in cognitive linguistics, this experiment extracted prefabricated chunks (prefabricatedchunks) from 2526 sentences. According to the semantic and grammatical features of these prefabricated chunks, a PrefabsLexicon containing 175,000 entries was constructed. To evaluate the lexicon' s effectiveness, the word segmentation experiment used a 56-minute dataset from another Lhasa Tibetan broadcaster as the test set. Compared with traditional Tibetan dictionaries, the PrefabsLexicon based on tone sandhi features achieved an F1-score of 0.92. Furthermore, in toneless Amdo Tibetan synthesis experiments, the synthesis quality MOS (Mean Opinion Score) improved to 4.17, indicating that the prefabricated chunk lexicon constructed based on tone sandhi features has cross-dialectal generalizability.

Full Text

A Prosodic Lexicon of Lhasa Tibetan: An Experimental Study Based on AI Speech Synthesis

LU Chen, LIU Chenning, ZHANG Xiao, ZU Yiqing*

Abstract This study proposes a method for constructing a prosodic lexicon for Lhasa Tibetan based on a continuous speech database, applicable to low-resource and complex languages. Using a small sample of high-quality data (3.77 hours, 2,526 sentences), the constructed prosodic lexicon significantly improves Lhasa Tibetan speech synthesis performance. The research reveals that tone sandhi in continuous Lhasa speech is constrained by semantic, morphological, and syntactic factors, reflecting the hierarchical organization and chunking rules of the language cognitive system. The phonetic manifestations of tone sandhi components include single-syllable citation tone, tone loss in single syllables, and two-syllable sandhi patterns. Which of these three patterns a syllable exhibits in connected speech is subject to three constraints: the first governed by word formation rules, the second by syntactic rules, and the third by high-frequency grammatical constructions. Inspired by the "Usage-Based Theory" in cognitive linguistics, this study extracts prefabricated chunks from 2,526 utterances. Based on the semantic and grammatical features of these chunks, we construct a Prefabs Lexicon containing 175,000 entries. To evaluate the lexicon' s effectiveness, segmentation experiments use a 56-minute dataset from another Lhasa broadcaster as the test set. Compared with conventional Tibetan dictionaries, the Prefabs Lexicon based on tone sandhi features achieves an F1-score of 0.92. Furthermore, in a toneless Amdo Tibetan synthesis experiment, the MOS (Mean Opinion Score) improved to 4.17, demonstrating the cross-dialectal applicability of the Prefabs Lexicon built on tone sandhi features.

[**Keywords**] Tone sandhi; Prefabricated chunks; Prosodic lexicon; Speech synthesis; Lhasa Tibetan

---

1. Introduction

In psycholinguistics, the mental lexicon studies lexical activities that language users engage in during everyday language comprehension and production. Jarema and Libben (2007:2) define the mental lexicon as "the cognitive system that constitutes the ability for conscious and unconscious lexical activities," emphasizing that the mental lexicon is the lexical activity itself—such as word comprehension—rather than the entity that enables it. Research in this area has had a direct impact on natural language processing (NLP). According to Faber and Mairal Usón' s (1999:20) study on English lexicons, NLP lexicons typically need to include: phonological knowledge of a language, lexical and expression structures, stress, and intonation; morphological information; syntactic configurations of lexical items in phrases and sentences; lexical

meanings and how these meanings combine to form sentence meaning; and pragmatic information such as communicative intentions. This demonstrates that a lexicon is not merely a list of words but must also contain multi-level information related to vocabulary—phonetic, morphological, syntactic—and knowledge of its dynamic changes in utterances. This means lexicon research cannot be separated from actual discourse and requires dynamic analysis in large amounts of continuous speech.

In speech science, synthetic speech is an important tool for verifying speech analysis results. As Kent and Read (1992:262) state, "only when we can reproduce a process do we truly understand it." Current speech synthesis experimental platforms based on sequence-to-sequence models can, on the basis of small amounts of high-quality continuous speech data, incorporate linguistic features at the textual, phonetic, and grammatical levels to build language models. This not only tests whether our phonetic analysis is correct but also helps verify whether certain linguistic concepts and knowledge systems are reasonable.

Text segmentation is a fundamental task in natural language processing. For Chinese and Tibetan, text segmentation typically refers to word segmentation. However, the concept of "word"does not naturally exist in these languages. From the perspective of natural text in both languages, there are no word boundaries —only syllable boundaries corresponding to one Chinese character or Tibetan syllable—reflecting native speakers' cognitive understanding of their language' s characteristics. This feature makes text segmentation a processing challenge for Chinese and Tibetan, with considerable insightful research and discussion. Overall, these approaches must balance multiple criteria: minimal independent usage units, semantic completeness, consistent grammatical properties, high frequency, pauses, and syllable count (Sun et al. 2001; Feng 2001; Wang 2001; Jiang 2003; Guan 2009, 2010; Ministry of Education 2015; Long and Liu 2016; GB/T 36452-2018 2018).

Currently, generative large language models like ChatGPT demonstrate remarkable natural language processing capabilities yet cannot stably complete simple word reversal tasks, frequently making errors. For instance, when GPT3.5 is asked to "reverse the word letter by letter: synthesis," it outputs "sisyhtnes." This problem is also evident in Chinese. When we ask GPT3.5 to "reverse this sentence character by character: you are a very smart robot," the output is "robot' s smart very one is you." Both tests show local fragments that fail to reverse correctly: "-yhtne-," "-robot-," and "-one-." The issue lies in the fact that natural language processing must first decompose text into minimal semantic units—tokenization—meaning the identification or segmentation into tokens. Tokens may be whole words, word fragments, Chinese characters, or words. Large models lack deep understanding of these linguistic units, conflating different levels of language units, whereas the human brain stores extremely rich linguistic units and related knowledge, enabling effortless manipulation of these units to complete language comprehension and generation.

Early linguistic theory held that due to limited memory capacity, the lexicon

should only contain unpredictable, most basic morphemes, while using rules to represent predictable structural information, thereby achieving a separation between lexicon and rules. The purpose was to avoid "redundant" information, thus excluding compound words from the lexicon. For example, Lieber (1980) and Selkirk (1984) likened the lexicon to a calculator containing a morpheme inventory and a rule system for combining basic morphemes into complex words. This "pocket calculator" language model is characterized by maximized computation and minimized storage, but it lacks a memory that stores computational processes and evaluates them, preventing it from learning from past experiences and steps (Baayen 2007:82).

However, the human brain clearly often acquires knowledge step-by-step from intermediate processes, gradually understanding and mastering complex matters, not always learning from minimal units and final outcomes. Consequently, Usage-Based Theory (J. L. Bybee and Beckner 2015) posits that specific learned instances in language and gradually emerging generalization patterns are stored together in memory. Speakers maintain rich memory representations, storing all details and rich experiences related to instances in actual representations. This study argues that the mental lexicon pre-stores language units at multiple levels and ranks, enabling us to simultaneously track and analyze multi-level linguistic information quickly during language use. In a neuroscientific study of speech comprehension, Ding et al. (2016) found that during language comprehension, neural activity in the human brain simultaneously tracks different levels of linguistic structure at different time scales. Based on Lhasa Tibetan speech synthesis experiments, we believe that the linguistic fragments involved in tone sandhi are key units in Lhasa language comprehension and use, and are core elements connecting lexical, syntactic, and prosodic levels. When building natural language processing lexicons, tone sandhi-related linguistic information must be explicitly expressed.

2. Lhasa Tibetan Speech Synthesis Experiment

The experimental sample (training data) consists of a self-built 3.77-hour, 2,526-sentence Lhasa Tibetan speech database recorded by a professional broadcaster. The study adopts a sequence-to-sequence speech synthesis method, performing direct encoding and decoding between input and output sequences. The speech synthesis model uses an autoregressive acoustic model and STRAIGHT vocoder, trained on 2,526 parallel data sentences totaling 3.77 hours at a 16kHz sampling rate. During training, 128 samples are input per step, predicting 4 frames at a time; during inference, a stepwise monotonic attention mechanism is used, with the traditional STRAIGHT vocoder reconstructing waveforms. The synthesis system uses the Wylie transliteration scheme (Wylie 1959), which corresponds one-to-one with Tibetan characters, as the input sequence. Experiments revealed that the choice of linguistic unit in text segmentation significantly affects final synthesis quality.

Early experiments segmented the 2,526 training sentences based on conventional

Tibetan dictionaries, with manual correction of automatic segmentation results. However, the resulting speech synthesis system achieved unsatisfactory MOS evaluation scores, with numerous phonological errors affecting sentence comprehension and unnatural rhythm. The root cause was the inability to reliably reproduce tone sandhi in natural speech—a widespread phonological phenomenon in Lhasa Tibetan and Chinese Wu and Min dialects, involving most linguistic components and directly affecting lexical semantic comprehension and prosodic naturalness in natural speech.

Every syllable in Lhasa Tibetan has a citation tone. When syllables enter words or sentences, tone sandhi occurs in two forms: single-syllable linguistic components lose their original tone and are weakened; two-syllable linguistic components change their citation tones, undergoing tone sandhi. Based on this phonological feature, we altered text segmentation rules, manually annotating the 2,526 sentences in the speech database. The segmentation units are no longer conventional dictionary words but three prosodic components: single-syllable citation tone components (marked as se1), single-syllable toneless components (marked as se0), and two-syllable tone sandhi components (marked as se2). We collectively refer to se1, se0, and se2 as SE units (sense elements) of Lhasa Tibetan. During data annotation, each sentence is segmented into a linear combination sequence of se1, se0, and se2. Consequently, the 2,526-sentence speech database has two comparable text segmentation methods: conventional dictionary word segmentation and SE unit segmentation. MOS comparison experiments showed that the SE-segmented synthesis system achieved significantly better results. The detailed experimental process is documented in "Basic Language Operating Units SE in Continuous Speech—Experimental Evidence from Tone Sandhi in Lhasa Tibetan" (Zu et al. 2022). The experimental results showed that synthesis based on conventional dictionary segmentation achieved a MOS of 3.45, while synthesis based on tone sandhi SE units achieved 4.25. We conducted objective error statistics on 50 synthesized sentences from this experiment. As shown in Figure 1 [Figure 1: see original paper], the 50 synthesized sentences using SE units not only showed significantly reduced tone sandhi errors but also improved accuracy in citation tone and initial/final production, demonstrating that linguistic unit settings affect not only higher-level prosodic features like pausing and rhythm but also lower-level phonetic features like segments and syllables learned by machine learning.

This experiment demonstrates that compared to conventional dictionary words, SE units extracted based on tone sandhi features may be dominant units underlying Lhasa Tibetan lexicon operation, better reflecting actual rules of phonological, morphological, and syntactic interaction during language use. The further research question is: what is the scope of tone sandhi implementation in continuous speech—how to segment tone sandhi domains in sentences, and what morphological and syntactic constraint rules govern this segmentation? From an engineering perspective, since the MOS experiment has proven that segmenting text into SE units yields better results, the next question is: how to achieve automatic segmentation of SE units, and what morphological and syntactic

rules are needed to ensure correct automatic segmentation? To address these questions, this study analyzes and annotates tone sandhi, morphology, and syntax in the 2,526-sentence speech database, summarizes the morphological and syntactic constraints for tone sandhi phenomena, and divides the analysis and prediction of tone sandhi components in speech synthesis work into two levels: lexicon and syntax. Based on this, we construct a prosodic lexicon of 175,000 entries and demonstrate its effectiveness in improving speech synthesis through two experiments.

3. Constraints on Lhasa Tibetan Tone Sandhi

3.1 Tone Sandhi Patterns and Domains

Tone sandhi research typically involves two tasks: tone sandhi patterns and tone sandhi domains. We must first understand tone sandhi patterns—the possible tonal patterns when two syllables undergo tone sandhi. However, to deeply investigate the scope of tone sandhi implementation, i.e., tone sandhi domains, we must consider more constraints: under what conditions sentence components undergo tone sandhi. Studying these constraints helps us better understand the application rules of tone sandhi in language, which is crucial for building machine learning models that automatically segment tone sandhi domains and predict tone sandhi. Therefore, our work aims to explore these constraints to accurately determine tone sandhi domains when analyzing continuous speech.

Previous Lhasa Tibetan tone sandhi research has thoroughly analyzed tonal patterns (Qu 1981b; Hu et al. 1982; Zhou 1983; Yu 1983; Xu 2015). Modern Lhasa has four citation tones, which we annotate as H, R, L, F (high, rise, low, fall). All syllable (Tibetan character) phonological information must be stored in the lexicon. When two syllables undergo tone sandhi under morphological and syntactic constraints, there are five tone sandhi patterns: HH, HF, LR, LF, LH. Which pattern two syllables exhibit depends on the initial syllable' s onset type and the final syllable' s rhyme type. Based on Tibetan orthography, a syllable' s onset is divided into high and low categories, and rhymes are divided into three types:

1. Rhymes with consonant letters l, r, m, n, ng, etc., are smooth rhymes;
2. Rhymes with consonant letters s, d, ms, ngs, b, bs, g, gs, etc., are checked rhymes;
3. Rhymes without consonant letters are open rhymes.

Tone sandhi derivation rules are shown in Table 1 .

Previous Lhasa studies mostly analyzed tone sandhi patterns when two syllables were known to undergo sandhi, focusing on phonetic manifestations and summarizing five patterns: HH, HF, LR, LF, LH. However, few studies have used continuous speech databases to investigate the scope of tone sandhi rules in speech flow, i.e., tone sandhi domains. Our work, based on numerous sentences, investigates how machines can automatically segment tone sandhi domains in

text and predict tone sandhi performance in any sentence. This provides empirical evidence for inferring the operational levels and organizational patterns of tone sandhi units in the human brain.

In Lhasa Tibetan text, a Tibetan character may exhibit multiple sandhi forms in different linguistic environments. For example, "tshod lta" may be a noun meaning "experiment"or "pilot,"where the two syllables undergo sandhi with HH pattern. Alternatively, "tshod lta" may be a verb meaning "to try" or "to experiment," where the two syllables do not undergo sandhi and are read with citation tones F and H respectively. Similar to Chinese, the word "good" has different tones in the idioms "good deeds come in pairs" (hao3) and "busybody" (hao4) due to different semantic and grammatical structures (modifier-head vs. verb-object), requiring more comprehensive linguistic context for accurate tonal analysis. Therefore, although tone sandhi manifests at the lexical level, constraints on tone sandhi domains also involve phrasal and syntactic factors.

To ensure the synthesis system accurately expresses these Lhasa features, we annotated 2,526 audio and grammatical instances, summarizing tone sandhi performance across all linguistic fragments. In continuous Lhasa speech, there are single-syllable components that do not undergo sandhi, including numerous basic nouns, verbs, adjectives, pronouns, and adverbs. However, more numerous are components that do undergo sandhi: single syllables losing independent tone (weak pronunciation) and two-syllable sequences undergoing tone sandhi. We mark non-sandhi single-syllable fragments as se1, weakly pronounced single-syllable fragments as se0, and sandhi two-syllable fragments as se2, collectively called SE units. Segmenting 2,526 continuous utterances revealed that weak se0 and sandhi se2 fragments dominate actual language use.

Table 2 shows SE component statistics in 2,526 sentences. According to the statistics, 31,555 SE units appear in the 2,526 sentences, with 7,095 unique SE units after deduplication. Sandhi se2 and weak se0 account for 79.0% of the lexicon entries and 65.4% of occurrences in the 2,526 sentences. This indicates that Lhasa utterances primarily consist of sandhi components in actual use, a feature that must be fully reflected in language model development.

3.2 Three Constraints on Lhasa Tibetan Tone Sandhi

Based on annotation data from 2,526 sentences and related literature (Wang 1956; Hu 1980; Qu 1981a, 1981b; Hu et al. 1982; Tan 1982; Zhou 1983; Huang 1994; Qu and Jinsong 2000), we find that syllable sandhi mechanisms in connected speech are constrained by three conditions: the first governed by word formation, the second by syntax, and the third by high-frequency grammatical constructions. These three constraint types correspond to different sandhi manifestations.

Table 3 shows constraint factors for three types of tone sandhi components. The first type results from two-syllable word formation and can be automatically segmented through exhaustive lexical coverage. This includes: widespread

two-syllable sandhi phenomena in Lhasa, constrained by two-syllable compound word formation, corresponding to two-syllable content words. In 2,526 sentences, this sandhi component appears 5,243 times across 11,360 instances, making it the focus of continuous speech lexicon collection. Wang Zhijing (1994:23, 25) argues that Tibetan monosyllabic morphemes are more fundamental than words, yet discussing morphemes sometimes requires reference to words. From a diachronic lexical development perspective, Tibetan shows a trend from monosyllabic to disyllabic or polysyllabic words. For example: two morphemes form disyllabic content words through compounding, nya "fish" + khrab "scale, armor" → nya khrab "fish scale, armor" with LF pattern; two morphemes form new words through symbolic representation, ka (first Tibetan letter) + kha (second Tibetan letter) → ka kha "letter, Tibetan alphabet" with HH pattern; content morphemes plus derivational affixes form new words, re "hope (verb root)" + ba (affix) → re ba "hope (noun)" with LH pattern, etc.

The second type results from syntactic functions, including: monosyllabic function words losing citation tone. Tibetan postposed function words or clitics typically occur at chunk boundaries, which are often also prosodic boundaries. Due to articulatory mechanisms, components at chunk ends partially lose their original tonal contours. Postpositions retain lexical meaning, e.g., tang "and" has incomplete tonal shape but maintains original tone class, recoverable under emphasis. However, postposed clitics completely lose citation tone; these monosyllabic components are marked as se0. Though few in number, they have high usage frequency: 129 postposed clitics appear 8,853 times in 2,526 sentences, including case markers kyi, kyis, tu, nas; topic/pause markers ni; adversative clitics mos "although, however" ; and utterance-final particles like declarative so, ngo and interrogative lam, dam.

The third type consists of two-syllable structures: a grammatical marker plus a verb or adjective. Here, the grammatical marker does not become weakly pronounced. We argue that due to frequent collocation with verbs/adjectives forming a fixed construction, this combination pattern has high-frequency salience in language cognition. During tone sandhi, they do not follow the rules in Table 1 but form a special grammatical sandhi where the preceding syllable' s tone depends on the following syllable, similar to Mandarin third-tone sandhi. In 2,526 sentences, four types appear 271 times across 160 words: 1) [negative marker ma/mi] + [verb/adjective]; 2) [verb/adjective] + [postposed conjunction na]; 3) [verb] + [nominalization skabs, dus, phyir, etc.]; 4) [verb] + [imperfective aspectual affix gi/gyi/kyi]; [verb] + [continuous aspectual affix gin/gyin/kyin]; [verb] + [imperfective nominalization affix rgyu]. These special grammatical sandhi structures consist of a functional marker attached to a verb/adjective, with the distinctive feature that the verb/adjective' s part of speech is retained in the entire syntactic word, and these components are still marked as se2.

Bybee (2002) notes that frequently repeated sequences become more fluent as they automatize into wholes—prefabricated chunks (prefabs) with independent representations in memory that can be accessed and executed as units. Tra-

ditional information processing lexicons have obvious deficiencies in covering these prefabs. We argue that the third type of sandhi component belongs to prefabs formed by grammatical constructions and should still be included in the lexicon. For such components, we first create inventories of verbs, adjectives, and grammatical markers when building the prosodic lexicon, then encode the constraints for this sandhi component as rules in the lexicon.

4. Construction Principles of the Lhasa Tibetan Prosodic Lexicon

4.1 Homographs in Lhasa Tibetan

The prosodic lexicon serves two purposes. First, we aim to incorporate tone sandhi information into the lexicon to enrich linguistic knowledge for low-resource language synthesis models. Second, we need to resolve ambiguities of homographs with identical forms but different pronunciations and meanings during text-to-speech conversion. By evaluating the language model' s disambiguation capability, we can assess our understanding of human language systems and linguistic competence.

Discourse communication and text reading/writing are the most common language application scenarios. Writing and speech are the most reliable information sources for studying human language systems. Language is a symbolic representation system of thought, while writing is a secondary symbolic result of this system. Compared to machines, humans can seemingly effortlessly record complex language using one-dimensional linear writing symbols. When facing numerous polyphonic and polysemous words in text, humans can understand communicative intent, quickly resolve ambiguity, and produce correct pronunciation. This ability relies on a mature working network in our brains connecting writing symbols, semantic knowledge, grammatical knowledge, and phonetic signals. This capability represents human linguistic and cognitive competence and is the core issue of our research.

Tibetan is an alphabetic script but does not directly express tonal information, let alone tone sandhi. Due to Lhasa Tibetan' s lexical tones and tone sandhi, many homographs exist with identical writing but different semantics or grammatical functions. Native speakers can quickly read Tibetan character-based sentences and convert text symbols into correct speech streams. We argue that the cognitive basis for this behavior is the pre-storage of SE units reflecting tone sandhi information and their high-frequency combinations in the brain. The brain' s ability to quickly read numerous homographs results from the lexicon containing abundant prefabs and construction template information for forming these prefabs.

Tibetan script presents two disambiguation problems: first, monosyllabic Tibetan characters can represent different morphemes and grammatical functions through citation tone, tone change, and sandhi; second, two-syllable fragments

can distinguish word classes through sandhi or non-sandhi. For example, the Tibetan character ma appears 430 times in the Lhasa speech database with three pronunciation patterns corresponding to three grammatical functions, creating disambiguation challenges: ma as a noun "mother" is read with citation tone R; ma as a word-forming suffix, e.g., nyi ma "sun," undergoes regular sandhi with nyi, forming LH; ma as a preposed negative marker, e.g., ma thub "cannot," undergoes special grammatical sandhi with thub, forming HF.

Polyphonic characters only become stable in prefabricated structures and thus need holistic storage in the lexicon. Particularly, ma thub "cannot" represents a special grammatical sandhi that would not traditionally be included in the lexicon as a whole.

In research on Lhasa Tibetan homograph disambiguation, Laba Dunzhu et al. (2018) compiled 140 disyllabic polyphonic words, primarily in two forms: those ending in affixes ba and pa, and compound words not ending in ba and pa. By statistically analyzing word forms, part-of-speech tags, and tone sandhi information across 2,526 sentences, we obtained data for these two polyphonic word types, detailed in Table 4 .

The key to the first type is that ba and pa have both word-forming and morphological functions. As word-forming components, they tightly combine with preceding root morphemes, forming se2 sandhi patterns. When ba and pa function as perfective nominalization components for verbs, they become weakly pronounced (tone 0), while the verb morpheme retains its verbal part of speech and original tone, forming se1+se0 sandhi combinations.

The key to the second type is that two monosyllabic morphemes form nouns when undergoing sandhi but verbs when not. Similar cases exist in English, e.g., record and perfect, where stress position distinguishes part of speech and function in speech. In 2,526 sentences, 617 sentences contain these two polyphonic word types. Though few in number, they are widely distributed. MOS evaluation shows these words significantly impact overall sentence semantic comprehension.

Polyphonic word disambiguation is an important criterion for evaluating synthesis quality and effectively measures our linguistic analysis quality. As previously noted, Lhasa Tibetan' s difficulty primarily lies in sandhi component variations. Therefore, we must observe sandhi environments in continuous speech and supplement relevant linguistic knowledge to further enhance language understanding and analysis capabilities.

4.2 Prosodic Lexicon Construction Scheme

Cognitive psychology research suggests that we can only understand mental phenomena better when viewing them as organized, structured wholes. To reduce the number of items to process, we chunk them, bringing order and coherence to perception (Sternberg and Sternberg 2016). Directly storing numerous high-

frequency chunks can effectively reduce the computational burden on the brain' s language system, reserving cognitive resources for more important and novel analytical tasks. Beckner et al. (2009) note that language' s cognitive organization is directly built on linguistic experience, with frequently co-occurring components at phonological and syntactic levels gradually forming retrievable chunks in the language system, influencing online language processing.

Conventional NLP lexicons typically focus on morphemes, words, phrases, and named entities. However, we argue that lexical sequences formed by high-frequency reusable construction templates should also be included in the lexicon if they exhibit phonological features like sandhi or weakening, with their sandhi construction information recorded. For Lhasa Tibetan, besides specific se1, se2, and se0 units, we also include prefabs based on sandhi patterns of identical text fragments across different sentences, enabling the lexicon to contain more contrastive and disambiguation information.

Prefabricated chunks, also called lexical bundles or formulaic sequences, refer to fixed collocations of more than one word in language. Becker (1975) argues that actual speech generation primarily relies on previously known phrases, completed through repetition, modification, and concatenation. For communication and comprehension convenience, language' s main production mode is assembling previously heard text fragments. Kapatsinski and Radicke's (2009) English speech perception experiment also supports the prefab hypothesis, suggesting that words and ultra-high-frequency phrases are stored in the lexicon for access. Additionally, language acquisition research finds that children acquire language through prefabs, inducing construction rules and forming grammatical competence through repeated exposure and use, with these chunks stored holistically in the mental lexicon (Nattinger and DeCarrico 1992).

Usage-Based Theory posits that multi-word phrases can be stored in memory and enter the lexicon. Although some multi-word sequences have transparent semantics and forms, these prefabs provide typical combination patterns. From a language use perspective, there is no need to choose between storing unanalyzable units or compositional assembly, as speakers may have rich and diverse representations of sequences (Erman and Warren 2000; J. L. Bybee and Beckner 2015).

Prefabricated chunk templates are inevitably high-frequency reusable patterns in language. "High frequency" can refer either to the high usage frequency of fragments formed by these templates or to the templates' strong generative capacity, frequently invoked in language use. This study identifies three features of Lhasa Tibetan prefab templates:

First, prefab templates typically include tone sandhi information. For example, two-syllable sandhi fragments are prefabs. According to Table 2' s statistics, when dividing the 2,526 training sentences into citation-tone syllables, weak syllables, and sandhi disyllabic components, disyllabic sandhi fragments have the highest usage rate at 37.3%. The lexicon must comprehensively include

these fragments and their operational rules.

Second, when identical Tibetan characters exhibit different sandhi patterns across sentence structures, both the sandhi and non-sandhi structures involving these characters are prefab templates. From a language cognition perspective, the lexicon must include these contrasting structural patterns to facilitate rapid identification and differentiation of polyphonic words. Based on 2,526 annotated sentences, we can analyze and statistically observe the frequency, distribution, and phonological-grammatical performance of identical text fragments across training data, effectively observing dynamic changes of static linguistic fragments in actual language use. For example, in Table 4, when tshod lta functions as a verb meaning "to try, to experiment," both characters retain citation tones (se1+se1 structure). When functioning as a noun meaning "experiment, pilot," the two characters undergo sandhi as se2. Both structures are prefab templates that facilitate language comprehension efficiency. Furthermore, if tshod lta is followed by the verbal morpheme byed "do, perform," it forms a common trisyllabic verb structure tshod lta byed meaning "to test, to try," with sandhi structure se2+se1. This 2+1 trisyllabic verb is very common in modern colloquial Tibetan (Gesang 2004:394) and represents a prefab construction in language use. Storing such structures helps eliminate tone sandhi ambiguity.

Third, while grammatical constructions in linguistics can vary in size, the prefabs defined in this study are relatively compact structures in language, typically without internal pauses, approximating prosodic words in prosodic phonology.

In summary, the prefab templates mined from the 2,526-sentence speech database are shown in Table 5 . These prefabs and their structural templates are key to improving disambiguation levels and are components largely unrecorded in previous lexicons. For example, [verb] + [progressive aspect affix bzhin] is a se1+se0 structure. bzhin can be a noun "face, complexion" read as citation tone R (se1), or a progressive aspect marker attached to verbs expressing "in progress," read as weak se0. Since bzhin as a progressive marker necessarily co-occurs with verbs and is typically followed by utterance-final particles like ' dug or yod, we categorize [verb] + [progressive bzhin] + [' dug/yod] as a prefab, recording its SE structure se1+se0+se0 in the lexicon as disambiguation information. Since 2,526 sentences represent only a small Lhasa sample, we gradually accumulated 183 polyphonic words and 339 prefab structures through literature review, then added approximately 15,000 prefabs through large-text matching and manual screening.

The traditional Tibetan dictionary used in early research contained 160,000 entries. Supplementing these with SE information and adding the later 15,000 prefabs resulted in a final prosodic prefab lexicon of 175,000 entries containing prosodic information.

5. Experimental Validation

To validate the prosodic lexicon' s effectiveness, we conducted two experiments: F1-score model evaluation and MOS evaluation for Amdo Tibetan.

First, we used a new speech database and compared the accuracy of tone sandhi domain boundary prediction between the prefab prosodic lexicon and conventional dictionary segmentation using F1-score evaluation. Results showed that prosodic lexicon segmentation achieved higher accuracy in predicting tone sandhi boundaries than conventional dictionary segmentation, confirming the prosodic lexicon' s improvement effects on data from other speakers.

Specifically, we used the previously mentioned 2,526 annotated sentences as the training set to build two tone sandhi domain prediction models using the prefab lexicon and conventional dictionary respectively. We then used another Lhasa speaker' s 56-minute, 282-sentence speech dataset as the test set, with manual tone annotation as ground truth. Both models predicted boundaries for the unannotated 282 test sentences, with results compared to manual annotation using F1-score to evaluate match degree between machine prediction and human annotation (ground truth).

F1-score is a metric for binary classification model accuracy, combining precision and recall as their harmonic mean. Precision reflects the model' s accuracy—how many predicted positive samples are truly positive; recall reflects completeness —how many true positive samples are correctly predicted. This experiment treats tone sandhi domain boundaries as positive samples and non-boundaries as negative samples. Therefore:

TP: Number of correctly predicted tone sandhi boundaries
FP: Number of non-boundaries predicted as boundaries
FN: Number of boundaries predicted as non-boundaries

Experimental results show that using the conventional dictionary yields an F1-score of 0.84. With the prefab lexicon, accuracy reaches 0.894 and recall 0.937 on the 282-sentence test set, raising F1-score from 0.84 to 0.92. Generally, Chinese word segmentation achieves F1-scores around 0.95. Considering Tibetan' s widespread homograph phenomena and the scarcity of high-quality resources compared to Mandarin Chinese, 0.92 represents a commendable result.

Additionally, MOS evaluation experiments demonstrated the prosodic lexicon' s cross-dialectal transferability. Yixi Weisa · Acuo (2003, 2004) notes that although Amdo Tibetan lacks lexically distinctive tones, it has habitual pitch with disyllabic words showing "high-low" and "low-high" patterns reflecting nominal-predicative differences. Xu (2015) suggests that toned Tibetan exhibits "front-not-high, back-not-low"sandhi patterns, while toneless Tibetan shows "front-low, back-high" habitual pitch patterns, both possibly deriving from pre-tonal lexical pitch concomitants. During Lhasa and Amdo speech data analysis and annotation, we found significant consistency between Amdo' s stress-timing units and Lhasa' s sandhi units. Directly applying Lhasa' s prefab lexicon to Amdo

synthesis systems improved MOS scores to 4.17, indicating the prosodic lexicon' s universal linguistic value across Tibetan dialects.

6. Conclusion

Based on an AI speech synthesis experimental platform, this study analyzes the chunking and dynamic changes of tone sandhi units in Lhasa Tibetan continuous speech. Through experimental data analysis, annotation, and observation of synthesis effects, results support the "rich memory representations" viewpoint. Humans often solve problems through step-by-step thinking, hypothesis generation, and reasoning to reach final answers. In language acquisition, children obtain linguistic input by observing and perceiving their environment, gradually inferring grammatical rules and sentence structures to produce new sentences and expressions. This reasoning process involves gradual understanding and application of lexicon and grammar. We argue that besides containing basic atomic components (like characters or morphemes), the lexicon system also pre-stores prefabricated chunks containing rich prosodic and grammatical information. Lhasa Tibetan tone sandhi-related linguistic fragments are dominant units in the Tibetan lexicon, core elements connecting lexical, syntactic, and prosodic levels and reflecting language system operational rules.

References

[1] GB/T 36452-2018. 2018. Tibetan Word Segmentation Specification for Information Processing. State Administration for Market Regulation; Standardization Administration of China.

[2] Feng, Zhiwei. 2001. "Certain Non-grammatical Factors in Determining Segmentation Units." *Journal of Chinese Information Processing*, Issue 5.

[3] Gesang Jumian and Gesang Yangjing. 2004. *Practical Tibetan Grammar Tutorial (Revised Edition)*. Chengdu: Sichuan Nationalities Publishing House.

[4] Guan, Bai. 2009. "A Brief Analysis of Several Concepts in Tibetan Word Segmentation." *Journal of Tibet University (Natural Science Edition)*, Issue 1.

[5] Guan, Bai. 2010. "Research on Tibetan Segmentation Units for Information Processing." *Journal of Chinese Information Processing*, Issue 3.

[6] Hu, Tan, Qu Aitang, and Lin Lianhe. 1982. "Experimental Study on Tones of Tibetan (Lhasa Dialect)." *Language Research*, Issue 1.

[7] Hu, Tan. 1980. "Research on Tones of Tibetan (Lhasa Dialect)." *Minority Languages of China*, Issue 1.

[8] Huang, Bufan. 1994. "Conditions for the Emergence and Differentiation of Tones in Tibetan Dialects." *Minority Languages of China*, Issue 3.

[9] Jiang, Di. 2003. "Methods and Process of Chunk-based Segmentation in Modern Tibetan." *Minority Languages of China*, Issue 4.

[10] Language Information Management Department, Ministry of Education (Ed.). 2015. *Tibetan Latin Transcription Scheme (Draft), Modern Tibetan Word Segmentation Specification for Information Processing (Draft), Modern Tibetan Part-of-Speech Tagging Set Specification for Information Processing (Draft).* Beijing: Commercial Press.

[11] Laba Dunzhu, Ou Zhu, Zu Yiqing, and Pei Chunbao. 2018. "Research on Disambiguation Methods for Tibetan Homographs with Same Form but Different Pronunciations." *Journal of Chinese Information Processing*, Issue 7.

[12] Long, Congjun and Liu Huidan. 2016. *Research on Theories and Methods of Tibetan Automatic Word Segmentation.* Beijing: Intellectual Property Publishing House.

[13] Qu, Aitang and Jinsong. 2000. *Theories and Methods of Sino-Tibetan Language Research.* Beijing: China Tibetology Publishing House.

[14] Qu, Aitang. 1981a. "Tones of Tibetan and Their Development." *Language Research*, Issue 1.

[15] Qu, Aitang. 1981b. "Tone Sandhi in Tibetan." *Minority Languages of China*, Issue 4.

[16] Sun, Maosong, Wang Hongjun, Li Xingjian, Fu Li, Huang Changning, Chen Songcen, Xie Zili, and Zhang Weiguo. 2001. "Word List for Modern Chinese Segmentation for Information Processing." *Applied Linguistics*, Issue 4.

[17] Tan, Kerang. 1982. "A Preliminary Discussion on Tone Classification and Notation of Lhasa Tibetan." *Minority Languages of China*, Issue 3.

[18] Wang, Yao. 1956. "Tones of Tibetan." *Chinese Language*, Issue 6.

[19] Wang, Hongjun. 2001. "Internal Structure of the 'Word List for Modern Chinese Segmentation for Information Processing'and Structural Characteristics of Chinese." *Applied Linguistics*, Issue 4.

[20] Wang, Zhijing. 1994. *Grammar of Lhasa Spoken Tibetan.* Beijing: Minzu University of China Press.

[21] Xu, Shiliang. 2015. "Habitual Pitch in Toneless Tibetan and Tone Sandhi in Tonal Tibetan." *Language Research*, Issue 4.

[22] Yixi Weisa · Acuo. 2003. "Mixing of Tibetan and Chinese Languages in 'Daohua' and Research on Deep Language Contact." PhD Dissertation, Nankai University.

[23] Yixi Weisa · Acuo. 2004. *A Study of Daohua.* Beijing: Nationalities Publishing House.

[24] Yu, Daoquan. 1983. *Tibetan-Chinese Lhasa Colloquial Dictionary*. Beijing: Nationalities Publishing House.

[25] Zhou, Jiwen. 1983. *Tibetan Phonetic Teaching Materials (Lhasa Pronunciation)*. Beijing: Nationalities Publishing House.

[26] Zu, Yiqing, Lu Chen, Ou Zhu, Zhu Ronghua, Liu Chenning, Shao Pengfei, Lubuta, Zhang Xiao, and Hu Guoping. 2022. "Basic Language Operating Units SE in Continuous Speech—Experimental Evidence from Tone Sandhi in Lhasa Tibetan." *Contemporary Linguistics*, Issue 4.

[27] Baayen, R. H. 2007. 5: Storage and computation in the mental lexicon. In *The mental lexicon* (pp. 81–104). Brill.

[28] Becker, J. D. 1975. The phrasal lexicon. In *Theoretical issues in natural language processing*.

[29] Beckner, Clay, et al. Language is a Complex Adaptive System: Position Paper. *Language Learning*, 2009, 59(s1): 1–26.

[30] Bybee, J. Sequentiality as the Basis of Constituent Structure. In T. Givón & B. F. Malle (eds.), *The Evolution of Language Out of Pre-Language*. Amsterdam: Benjamins, 2002: 109–132.

[31] Bybee, J. L., and Beckner, C. 2015. Usage-Based Theory. In B. Heine and H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

[32] Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1), 158–164.

[33] Erman, B., & Warren, B. 2000. The idiom principle and the open choice principle. *Text & Talk*, 20(1), 29-62.

[34] Faber, P. B., and Mairal Usón, R. 1999. *Constructing a lexicon of English verbs*. Berlin; New York: Mouton de Gruyter.

[35] Jarema, G., and Libben, G. 2007. 1: Introduction: Matters Of Definition And Core Perspectives. In *The mental lexicon* (pp. 1–6). Brill.

[36] Kapatsinski, V., & Radicke, J. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. *Formulaic language*, 2, 499-520.

[37] Kent, R. D., and Read, C. 1992. *The Acoustic Analysis of Speech*. Singular Publishing Group.

[38] Lieber, R. 1980. *On the organization of the lexicon*. (PhD Thesis). Massachusetts Institute of Technology.

[39] Nattinger, J. R., & DeCarrico, J. S. 1992. *Lexical phrases and language teaching*. Oxford University Press.

[40] Selkirk, E. 1984. *Phonology and Syntax- The Relation between Sound and Structure* (MA). Cambridge, MIT Press.

[41] Sternberg R J, Sternberg K, Mio J. 2016. *Cognitive psychology.* Cengage Learning Press.

[42] Wylie, T. 1959. A standard system of Tibetan transcription. *Harvard Journal of Asiatic Studies*, 22, 261–267.

Correspondence:

LU Chen (First Author)
510632 Guangzhou Jinan University, School of Literature; No. 601 West Huangpu Avenue, Tianhe District, Guangzhou, Guangdong, China; 13001023722; yousiruan@qq.com; WeChat: yousiruan

ZU Yiqing (Corresponding Author)
230088 Hefei iFLYTEK Co., Ltd.; Language Science Interdisciplinary Research Center, University of Science and Technology of China; Room 410, Block A, No. 789 Tianxi Road, Changning District, Shanghai (iFLYTEK Shanghai Technology Co., Ltd.); 13501684302; yqzu@iflytek.com; WeChat: wxid_{u7djzi0994mv22}

LIU Chenning
230088 Hefei iFLYTEK Co., Ltd.

ZHANG Xiao
230088 Hefei iFLYTEK Co., Ltd.

Note: Figure translations are in progress. See original paper for figures.