A Knowledge Representation Method to Implement A Taiwanese Tone Group Parser[1]

Yu-Chu Chang (張佑竹)

Abstract

A tone group parser could be one of the most important components of the Taiwanese text-to-speech system. In this paper, we offered the hypothesis of tonal government to emphasis the idea that if the allotone selection can be made for each word in a sentence then the tone groups will be separated within the sentence and supported our viewpoint with the implementation of a Taiwanese tone group parser. In addition to the description of using the symbol system to convert language expertise and heuristic knowledge into a knowledge base to cope with a frame-based corpus and a tone sandhi processor, the procedure of connecting the inference engine and the knowledge base to make allotone selection was also discussed. In the current version of the tone group parser, the average accuracy of inside test is 98.5%. The average accuracy of outside test is 94%. The experiment data of the study also reveals an important clue: the marking of the symbol system makes a higher contribution rate to the tone sandhi accuracy than the rule inference.

Keywords: Taiwanese, Tone Sandhi, Tone Group Parser, Knowledge Representation, Simulation

1. Introduction

Tone groups are the basic prosodic structure of Taiwanese[2]. A tone group parser is also one of the main components of a Taiwanese speech synthesis system (Liim, 2004; Tamura, 2010). This paper first explores the characteristics of Taiwanese from the perspectives of phonology and linguistic structure. It cross-validates the derivation process of how Taiwanese forms a unique tone group structure through tone sandhi with contemporary research on the phonology-syntax interface. It proposes the

[1] This article (in Chinese) was originally published in the International Journal of Computational Linguistics & Chinese Language Processing, Volume 22, Number 2, December 2017.

[2] Taiwanese originated from the Southern Min dialects of China. In 2006, the Ministry of Education announced the Taiwan Southern Min Romanization System, allowing Taiwanese to be publicly inherited in written form. The example sentences in this paper use the Ministry of Education's Taiwan Southern Min (hereinafter referred to as Taiwanese) romanization with tone values marked. "Taiwanese" or "Taiwanese language" in cited papers is also translated collectively as Taiwanese.

Tonal-Government Hypothesis, arguing that determining the tones of words within a sentence beforehand can also determine the boundaries of tone groups. Subsequently, the paper explains a method for constructing a tone group parser by integrating knowledge representation techniques and word attribute analysis based on the theory of Taiwanese tone sandhi acquisition.

2. Literature Review

2.1 The Nature of Taiwanese from Phonological and Structural Perspectives

Taiwanese is a tone language. Tone sandhi refers to the phenomenon where the tone of a word changes in value due to the influence of adjacent words, a feature common in languages across China. What is relatively unique is that Taiwanese words exhibit a pervasive tone sandhi phenomenon. Every Taiwanese word has two tone forms: the lexical tone[3] and the sandhi tone. The final syllable of a word or tone group is pronounced with its lexical tone, while the remaining syllables are pronounced with their sandhi tones (Chiu, 1931; Wang Yu-de, 1955). Therefore, if and only if the last syllable of a word or group of words is read with its lexical tone, this word or phrase is a tone group. In other words, a Taiwanese sentence is a collection of tone groups. Tone groups are not only syntactic units that make up Taiwanese sentences but also complete semantic units and prosodic structures. Taiwanese is likely the only natural language that establishes tone group structures within sentences through tone sandhi (Chang, 2009).

2.1.1 Taiwanese Tone Sandhi and the Formation of Tone Group

There is a close relationship among tone, semantics, and syntax in Taiwanese. Listeners can distinguish different parts of speech (POS) and semantics from the same sentence based on tone. For example, "ke^{55}" (chicken/add), which has the same lexical tone, appears in examples (1) and (2) with different tone forms, parts of speech, and meanings.

(1) Tsit^{31} tsiah^{31} (sandhi tone) ke^{55} (lexical tone, noun) tsit^{54} kong^{55}-kin^{55}.

(Translation: This chicken weighs one kilogram.)

(2) Tsit^{31} tsiah^{31} (lexical tone) ke^{55} (sandhi tone, verb) tsit^{54} kong^{55}-kin^{55}.

(Translation: This one is one kilogram heavier.)

[3] The lexical tone is also called the juncture tone, and the sandhi tone is also called the context tone.

This example draws our attention to the fact that although the human brain can analyze the semantics and syntactic structure using the homophonic but differently written Chinese characters "雞" (chicken) and "加" (add) as well as the context, for a computer, the romanization in (1) and (2) is completely identical. Only after the autonomous semantic mapping is determined can the tone form of "ke^{55}" be established or syntactic analysis be performed. This part falls under the category of Strong AI and is a dilemma that dialogue systems must face.

Examples (3) and (4) illustrate that the tone form of the quantifier "tsit^{32} king^{55}" (this house/room) depends on the surrounding context.

(3) Tsit^{32} king^{55} (lexical tone) u^{33} tsai^{55} hue^{55}.

(Translation: This house has flowers planted.)

(4) Tsit^{32} king^{55} (sandhi tone) u^{33} tsai^{55} hue^{55} e^{23} tshu^{31} si^{33} guan^{53} tau^{55}.

(Translation: The house with flowers planted is my home.)

From examples (5), (6), (7), and (8), we can observe the changes in tone groups caused by inserting different forms (Chang, 2009). Regardless of the form inserted, every Taiwanese sentence ultimately forms a structure composed of combinations of tone groups.

(5) [A^{55}-bi^{53}] [beh^{32} khi^{31} Tai^{23}-pak^{32}.] (A-mi wants to go to Taipei.)

There are two tone groups[4].

After inserting the sandhi tone word "siunn^{33}" (want/think), the number of tone groups remains unchanged.

(6) [A^{55}-bi^{53}] [siunn^{33} beh^{32} khi^{31} Tai^{23}-pak^{32}.] (A-mi wants to go to Taipei.)

After inserting the lexical tone word "pai^{31}-it^{32}" (Monday), the number of tone groups increases to three.

(7) [A^{55}-bi^{53}] [pai^{31}-it^{32}] [beh^{32} khi^{31} Tai^{23}-pak^{32}.] (A-mi is going to Taipei on Monday.)

[4] The symbol [ ] denotes the boundary of a tone group

After inserting the tone group "tse^{33} gu^{23}-tshia^{55}" (ride a bullock cart), the number of tone groups increases to three.

(8) [A^{55}-bi^{53}] [beh^{32} tse^{33} gu^{23}-tshia^{55}] [khi^{31} Tai^{23}-pak^{32}.] (A-mi wants to ride a bullock cart to Taipei.)

From a syntactic analysis perspective, a tone group must be an XP (X-phrase); however, not all XPs are tone groups. Tone groups may be predecessor structures that can be converted into XPs.

2.1.2 The Phonology-Syntax Interface

The Indirect Reference Hypothesis states that phonological rules are not directly influenced by syntax, but rather through prosodic structure as a medium linking phonology and syntax (Selkirk, 1986). This phenomenon is particularly evident in Taiwanese. Language learners apply prosodic cues to establish prosodic structures in order to acquire syntactic structure information. If children can mark the positions of tone groups in Taiwanese sentences, they can acquire useful syntactic-related knowledge (Tsay, 1999). Tone groups must be an important cue in Taiwanese language acquisition.

2.2 The Simulation of Language Acquisition

Regarding how language functions operate in specific areas of the cerebral cortex, Norman Geschwind pointed out that there are at least two regions in the cerebral cortex that have a significant impact on language abilities; these regions are precisely mapped to process speech information (Geschwind, 1979). Even though his theory that language ability primarily relies on the left hemisphere may be controversial, the human function of storing vocabulary in the brain's memory is unquestionable. Infants and young children learn language from speech perception. As they grow up, this perception mechanism remains intact (Eimas, 1985). Syllables, words, phrases, or prosodic units are likely stored in the memory blocks of the cerebral cortex. Therefore, we hypothesize that in the process of learning their mother tongue, Taiwanese speakers mark the part of speech and tone form of words in their memory.

Among Taiwanese words, some are pronounced with the sandhi tone, some with the lexical tone, and quite a few require determining the tone based on the part of speech and context. If a single rule is applied to roughly assign tones to a text—such that

monosyllabic words are read with the sandhi tone and polysyllabic words with the lexical tone—an accuracy rate of approximately 70% for tone sandhi can be achieved. In practice, the choice of a word's tone form is often related to the word's part of speech, adjacent words, and tone sandhi rules. Such words that require rule-based processing are also typically used by the speaker as tools to determine whether to extend the meaning; pronouncing a word with a sandhi tone indicates that the meaning referred to by the word is yet to be completed. A word read with the lexical tone is the boundary point of the tone group and the end of a complete semantic unit. Whether certain words should be read with the lexical or sandhi tone usually depends on the speaker, who must react in the instant before speaking.

The Prosodic Bootstrapping Hypothesis explains how children learn to use prosodic cues to help define tone groups, find syntactic structures, and acquire tone sandhi. This technique allows them to use the two forms of a word—lexical tone and sandhi tone—interchangeably (Tsay, 1999). A noteworthy fact is that native speakers of Taiwanese, even if they cannot detect the existence of tone sandhi rules or have never studied syntax seriously, can still accurately process tone sandhi and use the human brain's parser to identify tone groups. Similarly, when infants begin learning Taiwanese, they neither recognize words nor understand syntactic structures.

2.3 The Application of Knowledge Representation Methods

Marvin Minsky posited that problem-solving systems could be simulations of cognitive processes. After he proposed the application of frame theory (Minsky, 1975), knowledge representation became a focal technique in artificial intelligence research. Knowledge systems encompass three parts: a knowledge base, an inference engine, and a development interface. The knowledge base contains goals, rules, and domain-specific expert knowledge. The inference engine is responsible for the rule-inference process and strategy control (Chang, 1992). The development interface is used to communicate with users or link with other systems.

2.4 Current Research of the Taiwanese Tone Groups

Modern scholars noted the importance of studying Taiwanese tone groups in the mid-twentieth century (Wang Yu-de, 1955). Linguists have used syntactic analysis to identify Taiwanese tone groups (Cheng, 1968; Chen, 1987; Lin, 1994). Research has

sought evidence for tone sandhi as a prosodic boundary from phonetic experiments (Tsay, Myers, & Chen, 2000), or investigated tone group boundaries within the prosodic hierarchy by studying nasalization in Taiwanese, proposing that the tone group boundary is a prosodic unit of Taiwanese (Pan, 2003). Information engineering scholars have also attempted to build Taiwanese speech output systems (Liang et al., 2004) or applied part-of-speech (POS) tagging and tone sandhi rules to process Taiwanese tone sandhi (Iunn et al., 2007).

Pan (2003) pointed out that lexical tone words act as tone group boundaries, and that to determine the tones of words within a sentence, tone groups must be defined first. We, however, argue that if the tones of all words within a sentence can be determined, the tone groups can also be defined. Therefore, we propose the Tonal-Government Hypothesis to explain the relationship among Taiwanese tone groups, parts of speech, and tone forms.

3. Tonal-Government Hypothesis

Selkirk's Indirect Reference Hypothesis points out that syntactic structures do not directly constrain phonological rules, but rather affect phonological changes through prosodic structures as a medium. The relationship among them is: syntax -> prosodic structure -> phonology (Selkirk, 1986). However, since Taiwanese sentences are composed of tone groups, parts of speech and tone forms are highly likely to alter prosodic structures through tone sandhi rules. Examples (9) and (10) show that different tone forms of *khuann*^{31}, given the same part of speech, form different tone group structures and semantics. Examples (11) and (12) also show the regularity that the preceding word of a monosyllabic locative word is pronounced with a sandhi tone, whereas the preceding word of a polysyllabic locative word is pronounced with a lexical tone. The difference in the number of syllables between lai^{33} and lai^{33}-te^{53} affects the tone form and prosodic structure of the preceding word.

(9) [Li^{53} khuann^{31} (verb, pronounced with lexical tone)] [kam^{53} u^{33}]?

(Translation: Do you think there is/are?)

(10) [Li^{53} khuann^{31} (verb, pronounced with sandhi tone) kam^{53} u^{33}]?

(Translation: Can you understand it?)

(11) [Kong^{55}-hng^{23} (pronounced with sandhi tone) -lai^{33}] [u^{33} tsit^{54} tsiah^{32} kau^{23}].

(Translation: There is a monkey in the park.)

(12) [Kong^{55}-hng^{23} (pronounced with lexical tone)] [lai^{33}-te^{53}] [u^{33} tsit^{54} tsiah^{32} kau^{23}].

(Translation: There is a monkey inside the park.)

Although it is not the norm for phonology to directly affect syntactic structures in Taiwanese, the phenomenon of tone sandhi causing changes in prosodic structures is frequent. From Figure 1, it can be seen that in addition to prosodic structures affecting phonological changes, parts of speech, tone sandhi rules, and tone forms also affect or govern the formation of prosodic structures. It is worth noting that within the scope covered by the Tonal-Government Hypothesis, there is an obvious recursive phenomenon between the prosodic structure and the factors constituting phonological changes. The following sections will summarize the implementation method of the tone group parser to verify the possibility of the Tonal-Government Hypothesis.

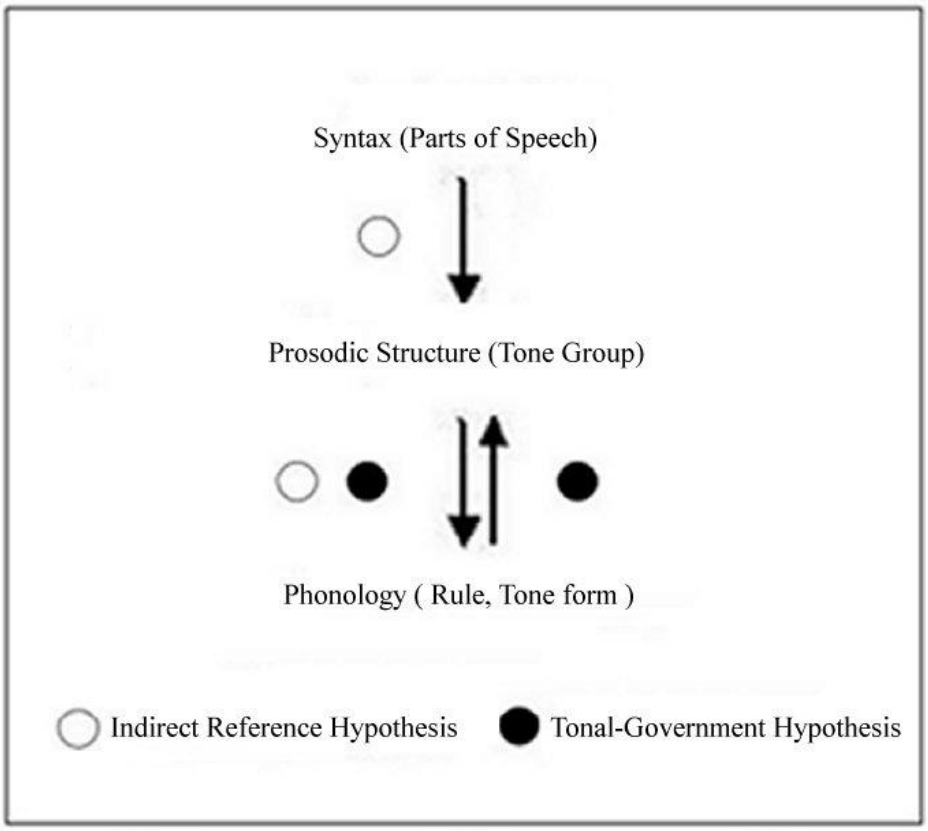

Figure 1. *The relationship among syntax, prosodic structure and phonology in Taiwanese with the related hypotheses*

4. The Method to Capture the Tone Groups from the Taiwanese Sentences

Tsay's discussion on Taiwanese tone sandhi acquisition highlights the importance of the feedback mechanism for simulation systems. Pan's experiments clearly indicate that tone groups are an important key to Taiwanese acquisition. These studies inspired our idea of building an artificial tone group parser using a personal computer. The inspiration for applying a symbolic system to find prosodic cues and using tone groups to resolve tone sandhi problems in words with multiple parts of speech comes from Selkirk's Indirect Reference Hypothesis.

Our concept is that once the words in a sentence are assigned the correct tone form, the words read with the lexical tone serve as the boundaries of the tone groups. The implementation method is based on theories such as Taiwanese tone sandhi acquisition, the Indirect Reference Hypothesis, and the Tonal-Government Hypothesis. It adopts a rule-inference strategy that primarily relies on default tone forms and secondarily on default parts of speech and the tone form patterns of preceding and succeeding words. After determining the tones of words within the sentence, the tone groups are extracted from the Taiwanese sentence.

Figure 2 is a schematic diagram of the basic architecture of the knowledge-based expert system within the Taiwanese tone group parser. The system is composed of a tone sandhi rule base, a frame corpus, and an inference engine. This expert system will be used to infer the tone value of each word in a sentence. Given that the generation of tone groups exhibits an obvious recursive phenomenon, the tone groups or preceding words of the tone groups extracted from the sentence by the system can also be fed back to the corpus through a recursive mechanism.

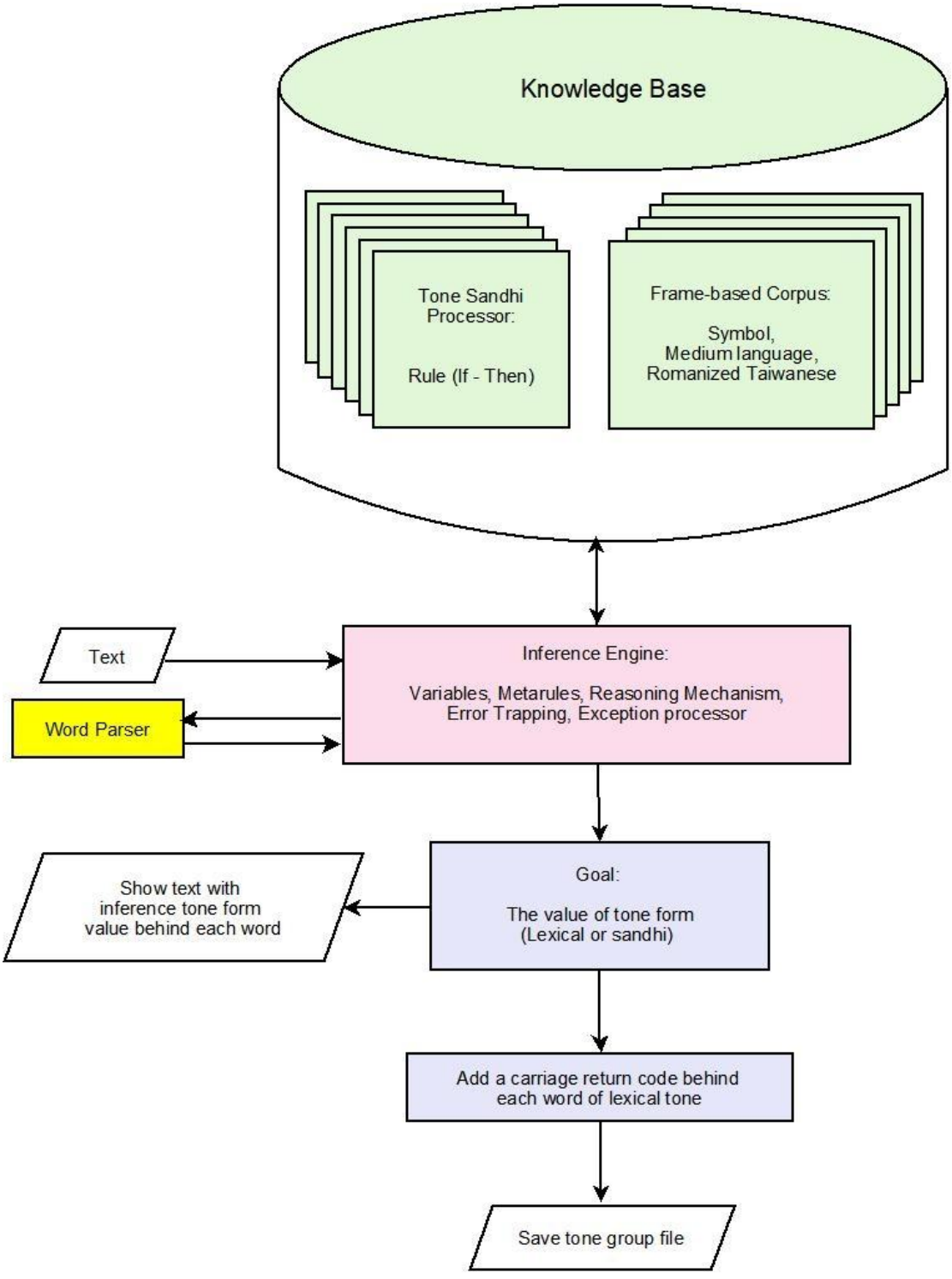

Figure 2. *Basic structure of the Taiwanese tone group parser*

5. The Schema of Implementation

The Tone group parser program is suitable for the Windows XP/Windows 7 operating systems[5] on personal computers. The implementation process is summarized as follows:

5.1 The Transformation of Linguistic Expertise and Heuristic Knowledge

Linguistic expertise and experience are primarily used for the rule base and corpus. After the required corpus is extracted from information such as dictionaries, professional literature, and fieldwork, Taiwanese language experts and knowledge engineers apply a symbolic system to tag the corpus and establish the tone sandhi rule base.

The design concept of the symbolic system originated from creativity during the early process of developing or modifying tone sandhi rules. The currently established symbolic system consists of three types of markers: Default mark of tone form, Default POS, and Mode. Each record in the corpus is assigned a set of symbols containing these three word attributes. This set of symbols is used to link the corpus and the tone sandhi processor.

When processing Taiwanese tone sandhi, with the help of the symbolic system and rule inference, even homonyms or words with multiple parts of speech can be assigned tones through the tone sandhi processing procedure. Once the corpus and the tone sandhi processor are constructed, the linguistic expertise is converted into a knowledge base. The following sections explain the three word attributes that make up the symbols and the inference process of using an inference engine to link the symbols, corpus, and tone sandhi processor.

5.1.1 The First Attribute: Default Mark of Tone Form

Taiwanese has many words that possess two parts of speech, or compound words formed by two parts of speech. Whether these words are pronounced with the lexical tone or the sandhi tone depends on the surrounding words and their relational positions within the sentence. Without rule inference through the tone sandhi

[5] The Windows XP/Win7 compatible version of the Taiwanese tone group parser can be downloaded from https://tonykelab.neocities.org/tycl

processing procedure, the tones cannot be determined. Marking the default tone form can screen words with a fixed tone form and eliminate unnecessary rule inference. The default tone markers and processing methods for Taiwanese words are shown in Table 1.

[Table 1. The list of default mark of tone form]

Default Tone Marker	**Default Tone Form**	**Applicable Words**	**Processing Method**
0	Fixed lexical tone	Words read only with lexical tone	No inference needed
1	Default sandhi tone	Monosyllabic words	Rule inference
2	Default lexical tone	Words read with default lexical tone such as phrases, neutral tone words, and foreign words	Rule inference
3	Fixed sandhi tone	Words read only with sandhi tone	No inference needed
#	Lexical or sandhi tone	Words that may be read with either lexical or sandhi tone	Rule inference
&	Fixed lexical tone	Tone groups or sets of tone groups	No inference needed

5.1.2 The Second Attribute: Default POS Mark

Taiwanese has a large number of compound words whose components are closely related to syntax and morphology. Tone sandhi causes changes in the tone of compound words, providing the speaker and listener with important information to distinguish different semantic or syntactic structures. However, compound words also make part-of-speech (POS) tagging more difficult in Taiwanese. Therefore, we use a loosely defined Default Part of Speech (DPOS) as the second attribute. The DPOS markers include n (noun/numeral), v (verb), a (adjective), c (conjunction), m (preposition), d (adverb), x (auxiliary verb), p (pronoun), u (quantifier), s (final particle), e (localizer), g (gerund), k (linking Verb), and & (tone group).

5.1.3 The Third Attribute: Mode Mark

Words or phrases with a default tone marker of 1, 2, or # require rule inference. Among these, words affected by preceding or succeeding words cannot have their tones determined solely by the default tone form and DPOS. It is necessary to use a set of mode markers applying two-level and three-level Boolean verification for rule inference.

- **Two-level mode markers** include a(-01), b(-11), c(-00), d(-10), e(10-), f(11-), g(00-), and h(01-).
- **Three-level mode markers** include j(000), k(010), m(101), n(111), p(001), q(011), r(100), and s(110).
- The marker x is used for words that do not require a mode marker.

Here, 0 represents the lexical tone, and 1 represents the sandhi tone. The - is used in two-level modes to indicate irrelevant preceding or succeeding word positions. For example, -01 represents a two-level mode in which the current word is read with the lexical tone, and the following word is read with the sandhi tone. The marker 101 represents a three-level mode in which the preceding and following words are read with the sandhi tone, while the current word is read with the lexical tone.

5.2 The Construction of a Frame-based Corpus

Existing natural languages, because they have been in use for a long time or have developed conventional usage, cannot be processed by computers using standard logical rules. Therefore, it is necessary to establish a computerized artificial language as a medium. The symbols and morphology of the medium language are not constrained by the natural language and can establish a mapping mechanism with the natural language. The advantage of using a medium language is that the syllable tone sandhi of words can be converted in advance without requiring further processing by the system.

Representing knowledge using an object-attribute-value model is a common method for constructing a corpus. In a frame-based corpus, a word or phrase can be regarded as an object. The default tone form describing the tone property is the attribute of the object, and the 2 assigned to the default tone marker is the value of the attribute. A hierarchical structure naturally forms between the object and its attributes. Therefore, the Taiwanese corpus can adopt the data structure of a general database and combine the default tone marker, default POS, and mode mark into a single set of symbols.

Each record has three fields: the symbol, the medium language string, and Romanized Taiwanese string with tone values marked by numbers. The fields are separated by a comma, such as "2nx,kangte,kang1-te7". The symbol marker consists of three letters; 2nx is used for a noun word or noun compound word that is read with the default lexical tone and does not require Boolean verification. The Romanized Taiwanese string can be a monosyllabic word, a polysyllabic word, a phrase, a tone group, or a clause. The medium language is the string corresponding to the Romanized Taiwanese. After all the corpus data undergoes word frequency statistics as well as attribute and functional analysis, it is sequentially stored in individual variable arrays similar to the human brain's long-term memory, allowing the system and rules to access it at any time.

Records are usually sorted by the number of syllables in the Romanized Taiwanese, with a higher number of syllables taking priority. Determining how to sort the records is practically quite difficult. Calculating the word frequency of common corpus data can serve as a reference for sorting, but the two are not exclusively correlated. Since Taiwanese morphology is not yet standardized, the rules for combining or separating syllables are highly complex. When designing the corpus search algorithm, a

morphological fault-tolerant mechanism must be considered to ensure the system operates smoothly during inference procedures.

5.3 The Design of the Tone Sandhi Processor

In practice, the tone sandhi processor performs rule inference based on the multiple attributes of words. All rules must be categorized in advance and prioritized. During categorization, they are divided into several main sections based on the relevant part of speech. The system first begins inference from the first section and, if necessary, moves to the next section to continue the inference of the target word. After the relevant rules are executed, the process can jump directly to the final section for actions such as debugging, Boolean verification, or concluding the inference. This inference procedure is known as forward chaining. The rules within all sections can be executed iteratively in a loop until the target words in the string array are processed one by one and the appropriate tone values are obtained.

Since some rules process the relationship between the preceding and following words, the inference process is inevitably subject to the mutual influence of different rules, which may alter the tone values already inferred for the adjacent words. The number of rules also affects the execution efficiency and tone sandhi accuracy of the parser. The more rules there are, the longer the inference takes, and the more significant the mutual interference becomes. Therefore, whenever the corpus or rule base is updated, an inside test must be performed to avoid compromising one aspect at the expense of another.

5.4 The Operating between Inference Engine and Knowledge Base

The inference engine is primarily used to access corpus information, and to initiate and control the tone sandhi processing procedure. The built-in search algorithm can match specified Romanized text sentences with words in the corpus to convert them into medium language sentences. These sentences are then segmented by a word parser into medium language strings and stored in an array.

Each medium language string can retrieve related attribute values from memory (variable arrays) and be assigned a target parameter, which is the default tone form

the system needs to infer. The parameter value assignment methods include using a default value, inferring the value through the system, or retrieving the value from the corpus. When the tone sandhi processor is called by the inference engine, the medium language strings in the array undergo the inference procedure in sequence.

The inference mechanism begins by matching the information provided by the corpus with the conditional parts of the various rules in the tone sandhi processor. When the IF part of a rule matches the relevant information, the command in the THEN part is executed. If no other rules in the tone sandhi processor are activated, the inference procedure terminates and the target inference is completed. Subsequently, the inference procedure for the next string begins until all medium language strings in the array have completed the inference operation, and the inferred target parameter values are returned to the inference engine.

The system may encounter a situation with insufficient information during inference; in this case, higher-priority metarules can directly debug, set, or change the relevant attribute values in the tone sandhi processor. The default metarules of the inference engine can also determine whether it is necessary to modify the inference results within the tone sandhi processor. This function is typically used to correct erroneous inference procedures or handle exceptional situations.

5.5 Semantic Identification, Fault Tolerance, and Machine Learning

Methods for semantic identification typically involve disambiguation or selecting from multiple meanings. In Taiwanese, some words can have their tones determined using general rules. For example, the preceding word of gah^{32} is always read with the sandhi tone. For homophones like ti^{33}, although they possess the multiple POS attributes of a noun (chopstick) and a preposition (at), their parts of speech can still be distinguished using relevant rules, and their tones can be assigned.

However, for homographs with different pronunciations and meanings, such as e^{23}e^{23} which can mean either “的鞋” (the shoes of) or “鞋的” (the soles/top of the shoe) with two different pronunciations and meanings, the system must have the capability to modify rule variables in order to differentiate them. When testing the tone sandhi processor, we formulated rules for the aforementioned phrases or clauses to let the machine judge the context and select the correct pronunciation, thus completing the tone assignment. The following output sentences are examples used by the tone group

parser to demonstrate Weak AI. The notation (1) represents the sandhi tone, and (2) represents the lexical tone.

(13) Tsit^{32} siang^{55} (1) e^{23} (2) e^{23} (1) e^{23}-bin^{33} (2) si^{33} (1) nng^{23}-a^{53}-phue^{23} (2) tso^{31}--e^{23} (2).

(Translation: The upper of this pair of shoes is made of split leather.)

(14) Tsit^{32} siang^{55} (1) e^{23}-bin^{33} (2) si^{33} (1) nng^{23}-a^{53}-phue^{23} (2) e^{23} (1) e^{23} (2) si^{33} (1) gua^{53}-e^{23} (2).

(Translation: The shoes whose upper is made of split leather are mine.)

Because the continuous or separated spelling of Taiwanese words often affects the inference results of tone sandhi, the establishment of a morphological fault-tolerant mechanism is also one of the key points in the design of the parser. As for how to enable the machine to learn and judge the context—whether through statistical probability analysis or direct rule guidance—it relies heavily on a large amount of knowledge, manpower, and computing resources, and can currently only undergo partial experiments. We cannot be certain whether children's acquisition of Taiwanese tone sandhi follows empirical rules entirely. However, the tone assignment results of (13) and (14) may demonstrate the feasibility of using tone sandhi rules to distinguish pronunciations and subsequently select semantics within the Taiwanese tone group parser.

5.6 Function Testing and the Result of Experiment

Once the words in a sentence are assigned their lexical tones, tone groups can be segmented. Therefore, reading a new text can generate new tone groups or preceding words of tone groups to serve as feedback units for the corpus. Since tone groups do not require inference to assign tones, the feedback mechanism can enhance the tone sandhi accuracy and execution efficiency of the parser. The more feedback the parser receives, the higher the tone sandhi accuracy becomes. Just like children learning a language, through the recursive feedback mechanism, the tone group parser can continuously evolve. In practice, this design can also be used to verify the discussions on tone groups by Tsay (1999) and Pan (2003).

This study involves two types of testing procedures: a debugging test designed for specific words or rules, and an accuracy and overall efficiency test conducted on the parser. The internal test corpus includes ten general articles and five sentences used to test specific words and rules. The external test corpus is derived from randomly extracted colloquial sentences from elementary school Taiwanese textbooks. In addition to speech judgments, the words in the tested texts are annotated with inferred tone values to calculate the tone sandhi accuracy rate. Currently, the average tone sandhi accuracy rate for internal testing is 98.5%, and for external testing, it is 94%.

During the program development period, we used the internal test corpus as the initial feedback material to update the knowledge base. When the average tone sandhi accuracy rate in internal testing approached 98.5% or began to converge, two knowledge base function experiments were conducted synchronously. The first test utilized only the symbolic system tags of the corpus without performing rule inference. The second test used no knowledge base at all, directly assigning the sandhi tone to monosyllabic words and the lexical tone to polysyllabic words. The following two formulas are used to calculate the contribution rate of rule inference and the contribution rate of symbolic system tagging. Both tests used the same internal test corpus.

Rule Inference Contribution Rate = Internal Test Accuracy Rate - First Test Accuracy Rate
Symbolic System Tagging Contribution Rate = First Test Accuracy Rate - Second Test Accuracy Rate

For general articles, the average tone sandhi accuracy rate of the first test was 91.41%, and that of the second test was 75.87%. The experimental results indicate that rule inference in the corpus contributes 7.09% to the tone sandhi accuracy rate, while symbolic system tagging contributes 15.54%. The relevant data are listed in Table 2.

For specific sentences, the average tone sandhi accuracy rate of the first test was 86.33%, and that of the second test was 60.35%. The experimental results indicate that rule inference in the corpus contributes 12.17% to the tone sandhi accuracy rate, while symbolic system tagging contributes 25.98%. The relevant data are listed in Table 3.

.

[Table 2. Tone sandhi experiment data for the general articles]

File No.	Test 1: Correct Words (A)	Test 1: Total Words (B)	Test 1: Accuracy	Test 2: Correct Words (C)	Test 2: Total Words (D)	Test 2: Accuracy
1	396	430	92.09%	343	430	79.77%
2	200	224	89.29%	157	224	70.09%
3	307	347	88.47%	254	347	73.20%
4	546	603	90.55%	454	603	75.29%
5	201	219	91.78%	153	219	69.86%
6	98	105	93.33%	68	105	64.76%
7	1006	1088	92.46%	869	1088	79.87%
8	607	669	90.73%	508	669	75.93%
9	178	203	87.68%	153	203	75.37%
10	613	654	93.73%	487	654	74.46%
Total	**4,152**	**4,542**	**91.41%**	**3,446**	**4,542**	**75.87%**

[Table 3. Tone sandhi experiment data for the special files]

File No.	Test 1: Correct Words (A)	Test 1: Total Words (B)	Test 1: Accuracy	Test 2: Correct Words (C)	Test 2: Total Words (D)	Test 2: Accuracy
11	411	489	84.05%	263	489	53.78%
12	663	773	85.77%	463	773	59.90%
13	606	667	90.85%	413	667	61.92%
14	456	556	82.01%	336	556	60.43%
15	675	771	87.55%	490	771	63.55%
Total	**2,811**	**3,256**	**86.33%**	**1,965**	**3,256**	**60.35%**

Because the proportion of words requiring rule inference in the specific sentences is higher than that in the general articles, the contribution rate of the symbolic system tagging and rule inference to the tone sandhi accuracy is also correspondingly higher, which is consistent with expectations.

Furthermore, the experimental data provides an important clue: whether for general articles or specific sentences, symbolic system tagging demonstrates a relatively higher contribution rate to tone sandhi accuracy than rule inference. In terms of the perception of Taiwanese tone groups, the word information stored in long-term memory may be more important and more efficient than the rules stored in short-term memory. The hypothesis that children learn to use prosodic cues from words to help define tone groups during the process of acquiring Taiwanese is also consistent with our experimental results.

6. Conclusion

In the process of learning their mother tongue, Taiwanese children can acquire knowledge of syntactic structures through tone groups. A reasonable assumption is that as more language knowledge accumulates in a child's mind, a highly efficient word tone sandhi processing mechanism is gradually constructed. The Taiwanese tone group parser can be said to be an experimental platform for artificial intelligence. We designed and improved a symbolic system as an important tool to convert linguistic expertise and experience into a knowledge base, which was used to construct a Taiwanese corpus and a tone sandhi processor, linking the simulation functions of tone sandhi acquisition with a speech output system and completing the tests.

The method of constructing a Taiwanese tone group parser using linguistic theories essentially establishes a simulation environment for tone sandhi acquisition using knowledge engineering techniques. The initial concept—that if all words in a sentence could be assigned tones, tone groups could be segmented from Taiwanese sentences—has been realized. This not only witnesses that AI development tools can help humans explore linguistic cognitive functions to understand the language acquisition process, but also presents the possibility of the Tonal-Government Hypothesis.

We attempted to process an unlimited number of sentences using a limited corpus and rules. However, constrained by the size of the knowledge base and computing resources, the current tone group parser cannot yet handle the tone assignment issues of autonomous semantics and some sentences. If this study can conduct supervised learning simulations or complete an automated feedback mechanism for tone groups and the preceding words of tone groups as a final solution to tone sandhi errors, it may improve the speech output functions of smart robots in the future.

References

Chang, T. Y. (1992). A multimedia-based bilingual instructional system using an expert system shell. (Unpublished master's thesis). University of Central Missouri, Warrensburg, MO.

Chang, Y. C. (2009). An introduction to Taiwanese Speech Notepad. Retrieved from ttps://archive.org/details/TaiwaneseSpeechNotepadenglishVersion.

Chen, M. Y. (1987). The syntax of Xiamen tone sandhi. Phonology, 4(1), 109-149.

Cheng, R. (1968). Tone sandhi in Taiwanese. Linguistics, 41, 19-42. Chiu, B. M. (1931). The phonetic structure and tone behaviour in Hagu (commonly known as the Amoy dialect) and their relation to certain questions in Chinese linguistics. T'oung Pao, 28(1), 245-342. doi: 10.1163/156853231X00105

Eimas, P. D. (1985). The perception of speech in early infancy. Scientific American, 252(1), 46-52.

Geschwind, N. (1979). Specialization of the human brain. Scientific American, 241(3), 180-199.

Iunn, U. G., Lau, K. G., Tan-Tenn, H. G., Lee, S. A., & Kao, C. Y. (2007). Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods. International Journal of Computational Linguistics and Chinese Language Processing, 12(4), 349-370.
Liang, M. S., Yang, R. C., Chiang, Y. C., Lyu, D. C., & Lyu, R. Y. (2004). A Taiwanese text-to-speech system with applications to language learning. In proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04), 91-95. doi: 10.1109/ICALT.2004.1357381

Lin, J. W. (1994). Lexical government and tone group formation in Xiamen Chinese. Phonology, 11(2), 237-276. doi: 10.1017/S0952675700001962

Liim, K. (2004). Medical Education and Research in Taiwanese Language Since 1990. Paper presented at the Symposium on Medical Taiwanese, Kaohsiung Medical University.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), The Psychology of Computer Vision. 211-277. New York, NY: McGraw-Hill.

Pan, H. H. (2003). Prosodic hierarchy and nasalization in Taiwanese. In Proceedings of the 15th ICPhS, 575-578.

Selkirk, E. O. (1986). Phonology and syntax: the relationship between sound and structure. Cambridge, MA: MIT press.

Tsay, J. (1999). Bootstrapping into Taiwanese tone sandhi. In Chinese Languages and Linguistics V, Symposium Series of the Institute of History and Philology, 2(5), 311-333. Taipei, Taiwan: Academia Sinica.

Tsay, J., Myers, J., & Chen, X. J. (2000). Tone sandhi as evidence for segmentation in Taiwanese. In Proceedings of the 30th Child Language Research Forum. 211-218.

王育德（1955）。台灣語の聲調。中國語學，41，3-11。[Ong, I.T. (1955). Taiwanese Tones. Journal of Chuugoku Gogaku, 41, 609-617.]

田村志津枝（2010）。初めて台湾語をパソコンに喋らせた男―母語を蘇らせる物語。 東京：現代書館。[Tamura, S.(2010).Hajimete Taiwango o pasokon ni shaberaseta otoko: bogo o yomigaeraseru monogatari.Tokyo, Japan: Gendai Shokan.]