Computational Linguistics and Chinese Language Processing

Vol. 4, No. 1, February 1999, pp. 59-84

©Computational Linguistics Society of R.O.C.

A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan

Chuan-Jie Lin* and Hsin-Hsi Chen*

Abstract

This paper presents a design of a Mandarin to Taiwanese Min Nan (abbreviated as Taiwanese hereafter) machine translation system. It is the first machine translation system which focuses on these two languages. An input Mandarin sentence is segmented, tagged and translated word by word according to the part of speech of each word. The candidates come from a Mandarin-Taiwanese dictionary. If more than one candidate exists, an example base is consulted. When a Mandarin word is not found in the Mandarin-Taiwanese dictionary, it is translated according to a Single-Character dictionary. The output can be in terms of either speech or text. For speech output, we also deal with the tone sandhi problem in changing the tone of each Taiwanese syllable. Because the mapping between Taiwanese syllables and Chinese characters is still a subject of disagreement, and the phonetic spelling coding systems are not familiar to everybody, speech output is useful but is also a challenge.

Keywords: Machine Translation, Mandarin, Speech Synthesis, Taiwanese, Min Nan, Tone Sandhi.

1. Introduction

Mandarin and Min Nan are two languages commonly used around world. According to Ethnologue [Grimes, 1996], the populations[1] of Mandarin- and Min Nan-speaking people are 885,000,000 and 49,000,000, respectively. They are ranked 1 and 21. In Taiwan, these two languages are also two of the major languages. This paper will study these two languages and present a Mandarin to Taiwanese Min Nan (abbreviated as Taiwanese hereafter) machine translation (MT) system, including a speech synthesizer of Taiwanese.

* Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. E-mail: cjlin@nlg2.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

[1] The population figures in [Grimes, 1996] refer to first language speakers in all countries.

Taiwanese has 7 tones (i.e., upper even, rising, upper departing, upper entering, lower even, lower departing, and lower entering), 17 initials, and 75 finals. The major problem with Taiwanese processing is its representation. Several methods have been proposed for representing Taiwanese syllables. One is to relate syllables to Chinese characters. But for some frequently used words, we cannot write down their corresponding characters without any doubt. The following shows a typical example:

beh (want to, will) " 我 **beh** 去學校 " **(I am going to school)**

Some people use " 要 " as its representation; some use " 欲 "; some even use " 卜 ". " 要 " and " 欲 " do represent the idea of "want to". However, " 要 " is read as "iau3" and " 欲 " is read as "iok8". It is hard to relate them to "beh". On the other hand, " 卜 " does have a colloquial reading of "beh", but it means "fortune-telling". That is, it is irrelevant to "want to". Another way to represent Taiwanese is to spell out the syllables. Similarly, there exist many coding systems, e.g., the International Phonetic Alphabet, the Missionary Romanization, and systems modified from the Chinese Phonetic Alphabet. Different from character representation, these coding systems can be transformed into one another. Therefore, they are equivalent. This paper will adopt the Missionary Romanization system. Table 1 [鄭良偉 , 1993] and Table 2 [中國大百科全書 , 1988] list the initials, the finals, and the tones of Taiwanese.

Table 1. *The Initials and the Finals of Taiwanese*

Initials	p, ph, b, m, t, th, l, n, k, kh, g, ng, h, ch, chh, s, j
Finals	a, ai, au, am, an, ang, e, eng, i, ia, iau, iam, ian, iang, io, iong, iu, im, in, ou, o, ong, oa, oai, oan, oe, u, ui, un, aN, aiN, eN, ouN, iN, iaN, iauN, iuN, oaN, oaiN, m, ng, ap, at, ak, ek, ok, iap, iat, iak, iok, ip, it, oat, ut, ah, auh, eh, ih, iah, iauh, ioh, iuh, ouh, oh, uh, oah, oeh, ahN, auhN, ehN, ihN, iahN, oaihN, mh, ngh

Table 2. *A Representation of the Tones of Taiwanese*

Tone number	Tone	Representation	Modified Representation
1	upper even	no mark, e.g. am	am
2	rising	´, e.g. ám	am2
3	upper departing	`, e.g. àm	am3
4	upper entering	no mark, e.g. ap	ap
5	lower even	^, e.g. âm	am5
6	rising		
7	lower departing	¯, e.g. ām	am7
8	lower entering	', e.g. a̍p	ap8

This paper is organized as follows. Section 2 presents the architecture of our Mandarin-Taiwanese machine translation system. Section 3 focuses on the lexical selection problems. Two strategies are proposed. Section 4 evaluates the performance of lexical selection and discusses it. Section 5 deals with speech synthesis. The tone sandhi phenomena are touched on. Appendix lists some translation results obtained using our Mandarin-Taiwanese machine translation system. Section 6 demonstrates an experimental system used on the Web. Section 7 offers concluding the remarks.

2. Architecture of a Mandarin-Taiwanese Machine Translation System

Many different approaches, e.g., rule-based [Bennett and Slocum, 1985], statistics-based [Brown et al., 1990], example-based [Nagao, 1984], and knowledge-based [Mitamura et al., 1991; Baker et al., 1994] ones, have been proposed for MT design. They have both advantages and disadvantages [Chen and Chen, 1995; Chen and Chen 1996]. No matter which kind of MT model is adopted, it has to capture the lexical and structural differences between the source and target languages. A typical transfer-based MT system is composed of a parser, a lexical transfer, a structural transfer and a generator. The parser analyzes the source sentences, and generates parsing trees. The lexical transfer selects the lexical items. The structural transfer captures the mapping of structures between source sentences and target sentences. The generator produces the target sentences.

Chao [1968] noted that the greatest degree of uniformity is found among all the dialects of the Chinese language in terms of grammar. Based on Chao's theory, we postulate that Mandarin and Taiwanese have similar structures. An example is shown as follows:

Mandarin:	他	今天	心情	很	好
Taiwanese:	伊	今仔日	心情	真	好

The above two sentences have the same word order. In our model, we focus on lexical selection. Source sentence analysis and the structure mapping between Mandarin and Taiwanese are neglected under the postulation. Figure 1 shows the overall architecture of our Text-to-Speech Mandarin-Taiwanese machine translation system.

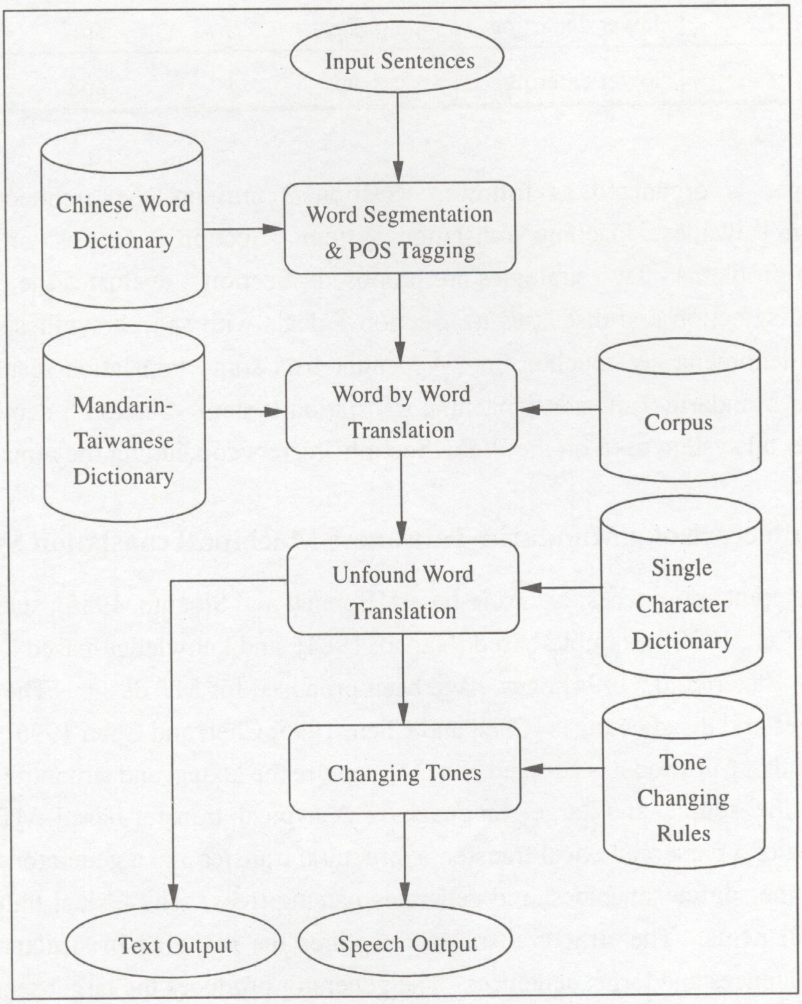


Figure 1 *Architecture of a Mandarin-Taiwanese Machine Translation System*

An input sentence in Mandarin is first segmented, then tagged, and finally translated word by word according to the part of speech of each word. The candidates come from a Mandarin-Taiwanese dictionary. If more than one candidate exists, an example base is consulted. When a Mandarin word is not found in the Mandarin-Taiwanese dictionary, it is translated according to a Single-Character dictionary. The output can be in terms of either speech or text. For the speech output, we also deal with the tone sandhi problem in changing the tone of each Taiwanese syllable.

3. Lexical Selection

3.1 Mandarin-Taiwanese Dictionary

A sentence in Mandarin cannot be translated into one in Taiwanese character by character. For example, we do not say 今天 (today) as "kin-thiN" (今天) in Taiwanese, but "kin-a2-jit8" (今仔日) instead. Thus, a word segmentation system is indispensable. We adopt the maximum matching criterion, 13 morphological rules selected from Lin's work [Lin, Chiang, and Su, 1993] and a lexicon trained from CKIP Corpus [CKIP, 1995] to identify word boundaries. The reason why we do not use the Mandarin-Taiwan dictionary for word segmentation is that the dictionary we adopt is not quite complete. Table 4 shows that 109 of the 1,000 most frequently used Mandarin words are not collected in this Mandarin-Taiwanese Dictionary. To get better segmentation and be easier integration with a proper noun identifier, we adopt a well-established word segmentation system.

In our experiments, the Taiwanese-Mandarin Dictionary (台華對譯詞庫 , 鄭良偉 , personal communication), developed by Professor Robert L. Chang, was employed in the lexical selection. There are 38,287 Mandarin words and 39,865 Taiwanese words. In total, there are 53,500 entries. An entry contains the corresponding Mandarin and Taiwanese words in Chinese characters, the phonetic spelling of the Taiwanese word, the POSes of these two words, the frequency of the Mandarin word, and some notes about the accent and other information.

The dictionary is organized in 8 columns, from A to H. Column A contains some information about Taiwanese words. For example, 'x' denotes that the Taiwanese word in this entry is rarely seen. Column B shows the phonetic spelling of the Taiwanese word using a modified Missionary Romanization system. Column C is the Chinese character representation of the Taiwanese word. Column D shows the corresponding Mandarin word, and column E its frequency. Columns F and G are the POSes of the Mandarin word and the Taiwanese word, respectively. The last column contains some other notes. For example, [2: 南] means that it is from the accent spoken in Southern Taiwan, while

[2: 北] refers to from the accent of Northern Taiwan. The notes in columns A, G and H are still incomplete. Therefore, we use only B, C, D, and F columns. The following examples illustrate the format of an entry:

B	C	D	F
{tek-pai5-a2%	{ 竹排仔 %	d% 竹筏 %	<Na
{sau3-se%	{ 掃梳 %	2d% 竹掃帚 %	<N>
{sau3-soe%	{ 掃梳 %	2d% 竹掃帚 %	<N>
{kui-nng5-lit8%	{ 規畚日 %	d% 一天到晚 %	<D

In the dictionary, a Mandarin word may have more than one Taiwanese translation. For example, the word " 會 " may be translated as "e7", "e7-tang3", "e7-hiau2", "e7-tit-thang", or "hoe7". Some are synonyms (e.g., "e7-tang3" and "e7-tit-thang"), and some are not. This is so-called lexical disambiguation problem. Table 3 shows the distribution of Mandarin words and their possible Taiwanese translations in the Mandarin-Taiwanese dictionary.

Table 3. *Translation Distribution of the Words in the Dictionary*

# of Candidates	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	24
Dictionary	0	29465	5445	1906	791	340	149	68	53	30	12	6	5	6	2	5	3	1
Top-1000 in CKIP	109	418	179	91	78	49	33	10	10	9	4	1	2	2	2	1	1	1

The first row depicts the corresponding number of Taiwanese candidates the Mandarin words have. For example, a total of 29,465 Mandarin words have only one Taiwanese translation each; 5,445 Mandarin words have two translations and so on. The second row shows similar statistics, but it considers part-of-speech information. When part-of-speech is not considered, 76.06% of the words have only one candidate, and a Mandarin word has 1.4 Taiwanese translations on average. At first glance, lexical selection seems to be easy. But we further compute another statistic. Here, we consider the 1,000 most frequently used Mandarin words, which were retrieved from Academia Sinica Balanced Corpus [CKIP, 1995]. The second row of Table 3 shows the distribution of these 1,000 words.

The average number of possible candidates increases to 2.49. If we further consider the frequency of these words used in the balanced corpus, i.e., if Formula (1) is used, the average number of candidates increases to 3.51. This number shows that lexical selection is not a trivial problem

$$\mathrm{Avg}_{1000} = \frac{\sum_{w \in Q} CD(w) \times f_{CKIP}(w)}{\sum_{w \in Q} f_{CKIP}(w)} \qquad (1)$$

where Q is the set of 1,000 most frequently used Mandarin words collected in the Mandarin-Taiwanese dictionary, $CD(w)$ denotes the number of candidates of w, and f_{CKIP}(w) is the number of occurrences of w in the CKIP Corpus.

3.2 Part of Speech

We adopt the following two criteria to select a suitable Taiwanese lexical item from the candidates.

(a) Part-of-speech;
(b) Corpus-based Method.

Table 3 shows that the most frequently used Mandarin words have 3.51 candidates in the Mandarin-Taiwanese dictionary on average. Thus, the first step is to reduce the number of candidates. Part of speech information is useful. For example, the word " 跟 " has several POSes: **Caa** (and), **VC** (to be with), **Na** (heel), *etc.* They correspond to different Taiwanese words. When it means "and" (**Caa**), its translation is "*kah*"; when it means "to be with" (**VC**), it is translated into "*toe3*"; and into "*kin*" when it means "the heel" (**Na**). In our dictionary, a column records the parts of speech of Mandarin words. We employ this information to reduce the number of candidates. Table 4 verifies this key point. The average number of candidates is reduced to 1.98 words when parts of speech are considered, and to 2.27 when Formula 1 is adopted. That is better than the word-information-only-method (2.49). We further proceed an investigation. Here, the top 1000 <Word, POS>-pairs are examined. Table 5 shows the statistics.

Table 4. *Translation Distribution of <Word, POS>-Pairs*

# of Candidates	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20
Dictionary	0	30964	5453	1806	714	295	117	58	38	18	9	3	4	2	1	4	1	1
Top-1000 in CKIP	1076	590	249	98	61	31	23	10	6	6	3	0	0	0	0	1	0	0

Table 5. The Number of Candidates of the 1,000 Most Frequently Used <Word, POS>-Pairs

# of Candidates	0	1	2	3	4	5	6	7	8	9	10	15
# of <Word,POS>-Pairs	239	402	186	69	44	24	16	9	3	4	3	1

The average number of candidates becomes 2.02 and 2.28, respectively.

Some translation rules are based on parts of speech. For example, the Mandarin word " 他 " has two meanings, i.e., he and his, and their corresponding Taiwanese words are different ("i" vs. "in"). The following examples clarify this point:

(1) 他 很 高興 (He is happy.)
i chin hoaN-hi2
(2) 他 爸爸 很 高興 (His father is happy.)
in a-pa3 chin hoaN-hi2

A translation rule is shown below:

```
IF Mandarin word=" 他 " and its POS="Nh", THEN
    IF the POS of the following word is "Na" or "Nc",
        THEN the corresponding Taiwanese is "in"
        ELSE the corresponding Taiwanese is "i"
```

3.3 Corpus-based Method

Part of speech information eliminates some of the candidates. However, it is still possible that more than one candidate has the same part of speech. In other words, we have to disambiguate further. Context information is necessary. The co-occurrence relationships between Taiwanese words are important. We can count how many times a word appears with other words and use the statistics to disambiguate the word usage. That is the so-called Markov model. However, a large scale of segmented Taiwanese corpus is needed to compute reliable statistics. Unfortunately, this is unavailable at present. We simplify the high-gram model to a unigram. That is, only the frequency of a word is employed. We always select high frequency words.

The statistics were computed based on the book " 國語常用虛詞及其台語對應詞釋例 " by Professor Robert L. Cheng [鄭良偉 , 1989]. This book originated from" 現代漢語八百詞 " [呂淑湘 , 1980], which lists frequently used Mandarin words and example sentences. Professor Cheng adds the most likely corresponding Taiwanese translations of each Mandarin word, and the corresponding Taiwanese sentences of the

Mandarin example sentences. Our approach is as follow: if a Taiwanese word appears in most of the example sentences of a Mandarin word, then the Taiwanese word is regarded as the correct translation of the Mandarin word. Of course, part-of-speech information is also considered. We segment and tag the Mandarin sentences first, and then extract the information we need. After the information is extracted, we rearrange the order of the entries in the dictionary. Since we prefer such a Taiwanese to be the translation of the Mandarin word with this part of speech, we simply set this Taiwanese word in the first entry of this Mandarin <Word, POS>-pair. During translation, we just choose the first entry of the matched <Word, POS>-pair.

3.4 Translation of Unfound Words

Table 4 shows that even the 1000 most frequently used words may not all be found in the Mandarin-Taiwanese dictionary, not to mention proper names such as those of persons, companies, and organizations. In order to deal with these unfound words, we have developed a Single-Character dictionary. It contains 5,680 single Chinese characters and their readings in Taiwanese. For example, " 資 " is read as "chu" in Taiwanese. These characters contain the 5,401 most frequently used characters as proposed by The Ministry of Education [教育部 , 1979] as well as those characters which may be used in Chinese names and transliterated foreign names [Chen and Lee, 1996]. The readings were selected from " 國台雙語辭典 " [楊青矗 , 1992]. Each character may have literary readings and colloquial readings. For Example,

	Literary Reading	Colloquial Reading
倩	chhian3	chhiaN3
水	sui2	chui2

4. Evaluation of Lexical Selection

4.1 Test Results

The test data came from the Evening News Broadcast of the Chinese Television Station (CTS) on October 16, 1996. The CTS Evening News was broadcast in both Mandarin and Taiwanese simultaneously. There were 789 Mandarin total sentences and 6,484 total words in our test data, and 758 Taiwanese sentences and 5,916 words in the Taiwanese version. A total of 31 Mandarin sentences did not have any corresponding Taiwanese sentences. Thus, these sentences were neglected in the tests.

Table 6. *Precision and Recall Rates in Lexical Selection*
(MAT: Number of Words Matched; R: Recall; P: Precision)

	Answer	Correctly Seg-Tagged Input				Plain Text Input			
	Set A	Set B	MAT	R	P	Set C	MAT	R	P
No POS	5916	6323	2790	47.16	44.12	6639	2619	44.27	39.45
POS	5916	6323	2738	46.28	43.30	6639	2443	41.29	36.80
POS + Corpus	5916	6323	3204	54.16	50.67	6639	2848	48.14	42.90

The Chang-Chou Accent and Chuan-Chou Accent are two of the accents of Taiwanese. In our experiment, we set the accent option to Chuan-Chou first. The experimental results are listed in Table 6. A denotes the correct answer in the test data. B and C denote the translation results of our system. To show the influence of a word segmentation system and a part-of-speech tagging system, B received input which was correctly segmented and tagged by hand, and C received input from our word segmentation and part-of-speech tagging systems. The first row depicts the precision rate and recall rate when POS was not considered. We randomly chose a translation from all the candidates. The results listed in the second row are those obtained when we employed POS information but randomly chose a translation from all the candidates of the same POS. The third row shows the results obtained when both POS and corpus-based knowledge were considered.

Table 6 shows that the correctness of the word segmentation and POS tagging systems indeed affected the lexical selection performance. The precisions of the segmentation and tagging systems were 88.0% and 78.69%, respectively. The performance of the segmentation system was not perfect, but this did not affect the translation performance so much as we expected. This is because that proper noun identification had not yet been integrated, and these proper nouns were not easily translated. This will be discussed later in Section 4.2.3.

Surprisingly, the recall rate and the precision rate dropped 1-3% when POS information was considered. In both cases B and C, integrating POS information and corpus-based knowledge worked better than did the other two methods. When we analyzed why the POS approach was a little worse than the No-POS approach, we found that the POS information of in our Mandarin-Taiwanese dictionary was incomplete. Thus, we did an investigation. This time, we focused only on those Mandarin words whose <Word, POS>-pairs appeared in the dictionary. Table 7 shows the results.

Table 7. *Examination of 4,609 Pairs Found In the Dictionary*

number of words which had no correct candidates	2837
number of words which had only one candidate	1641
number of words which had multiple candidates	1196
number of correctly selected translations in multiple candidates	1009
correct rate	84.36%

During translation, a total of 4,609 <Word, POS>-pairs were found in the Mandarin-Taiwanese dictionary, but only 2,837 appeared in the answer set. Among these 2,837 pairs, 1,641 had only one candidate (lexical selection is trivial in this case), and 1,196 had more than one candidate. Among these 1,196 pairs, 1,009 were correct, so the correct rate was 84.36%. This finding supports our point.

4.2 Discussion

Besides the propagation errors from segmentation and tagging, we found several other kinds of errors. They will be discussed in the following subsections.

4.2.1 Errors from the Test Data

Our test data were selected from the CTS Evening News. It is a live broadcast, so both Mandarin and Taiwanese broadcasters speak at the same time. It is interesting that the Taiwanese broadcaster always spoke more slowly and it took her a little more time to speak out the same sentence which Mandarin broadcaster said. Thus, the Taiwanese broadcaster tended to eliminate words and sentences to keep up with the Mandarin broadcaster. The following shows an example. The symbol (C) denotes the correct answer, and (D) denotes our translation result.

(3) 不過 行政 部門 對於 這個 建議 仍然 持 相當 保留 的 態度

(C) ia2-m7-ko3 heng5-cheng3 pou7-bun5 * * * * iu5-oan5 si7
chhi5-tioh8 siong-tong po2-liu5 e5 thai7-tou7

(D) m7-koh heng5-cheng3 pou7-bun5 tui3-u5 che ko3 kian3-gi7
iu5-goan5chhi5 siong-tong po2-liu5 tek thai7-tou7

The phrase " 對於這個建議 " did not appear in the Taiwanese sentence. That lowered the precision rate of our system. In contrast with the above situation, some phrases were added in the Taiwanese sentences but had no corresponding phrases in the Mandarin sentences. For example,

(4) 上午 （集合 在 台北市 世貿 廣場 前面） 接受 交通 局長 賀陳旦 的 檢閱

(C) cha2-khi2 chip8-hap8 ti7-leh tai5-pak-chhi7 se3-bou7 kong2-tiuN5 thau5-cheng5 chiap-siu7 kau-thong kiok8-tiuN2 ho7-tan5-tan3 e5 kiam2-iat8

(D) cha2-khi2 chiap-siu7 kau-thong kiok8-tiuN2 ho7 tin5 tan3 e5 kiam2-iat8

The phrase "集合在台北市世貿廣場前面" was eliminated by the Mandarin broadcaster. That lowered our recall rate.

Table 8. *Comparison between the Chang-Chou and the Chuan-Chou Accents*

CHUAN-CHOU	CHANG-CHOU	EXAMPLE (CHUAN-CHOU,CHANG-CHOU)
l	j	熱 (liat8, jiat8); 仁 (lin5, jin5)
e	oe	過 (ke3, koe3); 尾 (be2, boe2)
oe	e	買 (boe2, be2); 雞 (koe, ke)
iN	eN	平 (piN5, peN5); 姓 (siN3, seN3)
ui	oe	媒 (mui5, moe5); 血 (hui3, hoe3)
u	i	女 (lu2, li2); 箸 (tu7, ti7)
un	in	近 (kun7, kin7); 根 (kun, kin)

Furthermore, the accents also lowered the precision rate and recall rate. Recall that we adopted the Chuan-Chou accent in the preliminary test, but the reporter still read some words using the Chang-Chou accent. Table 8 shows comparisons between these two accents.

Another inconsistency was that a word could be spoken differently in the broadcast. For example,

(5) 總統府 資政 黃少谷 今天 凌晨 因為 感冒 引發 肺炎 病逝 在 台北 榮總

(C) chong2-thong2-hu2 chu-cheng3 ng5-siau3-kok kin-a2-jit8 thau3-cha2 in-ui7 kam2-mou7 in2-hoat hi3-iam7 ti7 tai5-pak eng5-chong2 koe3-sin

(6) 黃少谷 生 於 民國 前 十一年

(C) ng5-siau2-kok si7 ti7 bin5-kok cheng5 chap8-it-ni5 chhut-si3 ti7

The name " 黃少谷 " was read in two different ways. All the above problems show that a high quality bilingual evaluation is hard to prepare. In the experiments, we used a strict measure, so the test data influenced the precision rate and the recall rate severely.

4.2.2 Errors from Dictionary

Some errors resulted from dictionary. Consider the following example:

(7) 至於 診所 裏 的 醫生 護士 有 沒有 過失（責任）

(C) chi3-u5 chin2-sou2 * * i-seng hou7-su7 si7 m7-si7 u7 ke3-sit chek-jim7

(D) chi3-u5 chin2-sou2 lai7 e5 sian-siN hou7-su7 u7-but8-u7 ke3-sit

Here, " 醫生 " was translated as "sian-siN" in our system. The Taiwanese word "sian-siN" is an old-fashioned word, a respectful way to address a doctor. The inappropriate translation dragged down our performance.

4.2.3 Proper Noun Translation

Proper nouns are not easy to translate. Each character has literal reading and colloquial reading. It is not clear when to use which reading. The most famous case is " 大 ". " 林大海 " is read as "lim5-toa7-hai2", but " 大傑 旅行社 " is read as "tai7-kiat8 li2-heng5-sia7".

Table 9 shows our results for names of persons, locations, and companies. Most of the errors were due to incorrectly reading selection. Some of them were eliminated by the Taiwanese broadcaster (as the case discussed in Section 4.2.1). Other errors included mistakes in the Single Character Dictionary, accent, word-structure, error-reading or inconsistent-reading of the broadcaster. Location names were much easier to translate because many of them were collected in the bilingual dictionary.

Table 9. *Investigation on Proper Noun Translation*

Names of	Correct	Incorrect		
		Selection	Elimination	Others
Persons	67	45	17	38
Locations	147	15	13	10
Companies	28	16	2	0

4.2.4 Quantifier Problem in Translation

The translation of quantifiers is a common problem in machine translation. So far, we have not dealt with this problem. The following shows an example:

(8) Mandarin: 一 把 雨傘
Taiwanese: 一 支 雨傘
chit8 ki hou7-soaN3

The quantifier for an umbrella in Taiwanese is " 支 " instead of " 把 ". But " 一 把 稻草 " is still translated as " 一 把 稻草 ".

4.2.5 Other Problems

Word-by-word Translation is not always acceptable. Consider the following two examples.

(9) 逮捕 _VC 了 _Di 南投 _Nc 地區 _Nc 首惡 _Na 份子 _Na
(C) liah8 tioh8 lam5-tau5 te7-khu siu2-ok hun7-chu2
(D) liah8 liau2 lam5-tau5 te7-khu siu2-ok le2

(10) 並且 _Cbb 將 _P 四 _Neu 個 _Nf 人 _Na 同時 _Nd 逮捕 _VC
(C) ji7-chhiaN2 chiong si3 e5 lang5 tong5-si5 liah8=khi0-lai0
(D) ji7-chhiaN2 chiong su3 ko3-jin5 siang5-si5 liah8

In example (9), " 逮捕 " is translated as "liah8". In example (10), " 逮捕 " should be translated as " liah8=khi0-lai0". This is because if we place "liah8" after the target, we have to say, e.g., " 逮捕 他 " (to arrest him), "chiong i liah8=khi0-lai0" (將 他 逮捕) or "kah8 i liah8=khi0-lai0" (把 他 逮捕).

5. Speech Synthesis

5.1 Tone Sandhi

Tone sandhi is a common problem in tonal languages like Chinese. In Mandarin, when two or more third-tone characters are put together, the preceding third-tone characters is read using the second tone. In Taiwanese, there is also a tone sandhi problem. However, it is quite different from the one in Mandarin.

The tone used to read a character individually is called its original tone. When more than one character is spoken, we always change the tones of the preceding characters. Consider the following example:

(11)	Mandarin:	他	今天	心情	很	好
	Taiwanese:	伊	今仔日	心情	真	好
	Original tones	i	kin-a2-jit8	sim-cheng5	chin	ho2
	Changed tones	i7	kin7 a jit8	sim7 cheng5	chin7	ho2
		he	today	mood	very	is-good

In the above example, the last characters of " 心情 " and " 今仔日 " retain their original tones. The word " 好 " is at the end of the sentence, so its tone remains unchanged, too. The tones of the other characters are changed.

In the Mandarin-Taiwanese dictionary, the phonetic spelling of a word follows the original tones. During speech synthesis, we have to change the tones in order to speak fluently. When and how the tones should be changed depends on situation and accent. Section 5.2 will discuss this topic in detail.

5.2 The Rules

5.2.1 General Rules

Figure 2 shows general rules for the Chang-Chou accent and Chuan-Chou accent [鄭良偉 , 1993]. The tone numbers are shown in Table 2. The symbols p, t, and k are stop endings, and h is a glottal stop ending. The rules for the Chang-Chou accent and Chuan-Chou only differ a little. In the Chang-Chou accent, the lower even tone (the 5th tone) changes to the lower departing tone (the 7th tone). In contrast, the lower even tone (the 5th tone) changes to the upper departing tone (the 3rd tone) in the Chuan-Chou accent.

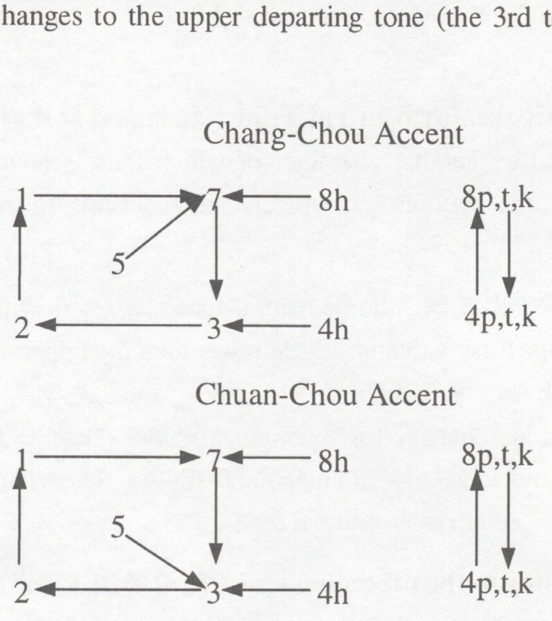


Figure 2 *Tone-Changing Rules*

In our study, the upper entering tone did not change to the lower entering tone. The changed tone was most like a rising tone except that it was a checked sound while the lower entering tone shared the same tone value with the lower departing tone. For better speech output, we added another tone number 9, to indicate that the tone changed from the upper entering tone. The cause might have been the accents again because in some accents their tones 8 and 9 (defined by us) are the same (i.e., the 8th tones in different accents are different).

5.2.2 Tone-Changing Rules before "a2" (仔前變調規則)

The character "a2" (仔) is a special character in Taiwanese. It is a suffix and is always combined with other characters to form a word. For example,

(12) 椅 "i2" 椅仔 "i2-a2" " i a2"
(13) 歌戲 "koa-hi3" 歌仔戲 "koa-a2-hi3" "koa7 a hi3"

The tone of the character before "a2" also changes, but the rules are different from the general rules mentioned in the previous section. We adopted the analyses [楊秀芳, 1991] shown in Table 10.

Table 10. *Tone-Changing Rules before "a2"*

Original	1	2	3	4	5	7	8
Chang-Chou	7	2	2	9 (ptk) 2 (h)	7	7	8 (ptk) 7 (h)
Chuan-Chou	7	2	2	9 (ptk) 2 (h)	3	7	8 (ptk) 7 (h)

5.2.3 Tone-Changing Rules for Neutral Tone (輕聲調變調規則)

If a character is followed by another character of which tone is neutral, it retains its original tone. However, the neutral tone will change according to the following two tone-changing rules [洪維仁, 1996].

(a) **Solid tone value** (固定調): Some neutral-tone characters will retain a solid tone value. This tone value is a little lower than the upper departing tone. Characters such as " 來 " (lai0), " 去 " (khi0), and " 先生 " (sian0-siN0) determine the tone-change, for example, " 過來 " ("koe3=lai0"), " 無去 " ("bo5=khi0"), and " 陳先生 " ("tan5=sian0-siN0"). The characters " 過 ", " 無 " and " 陳 " retain their original tones.

(b) **Change according to the preceding tone** (隨前變調): The tones of some neutral-tone characters will change according to the tones preceding them. The rules are shown below:

```
if the preceding tone is the upper even tone,
    then the tone of the neutral-tone character change to the upper even
        tone
    else if the preceding tone is one of the upper tones,
        then it change to the upper departing tone
        else it change to the lower departing tone.
```

的 "e5", 矣 "ah0" and the pronouns are such characters. Examples are:

(14) 新的 "sin=e5" "sin e"
(15) 予我 "hou7=goa2" "hou7 goa7"

However, if a neutral tone character of this kind is not in the end of a syntactic structure, or if, as in (15), the pronoun is indeed combined with the following word the tone will change according to the general tone-changing rules.

5.2.4 Triple-Character Adjectives (三疊形容詞變調規則)

In Taiwanese, we sometimes will triple a one-character adjective (e.g., sour, " 酸 ") to emphasize the situation (in this case, " 酸酸酸 "). This makes the sound more interesting. The tone of the second character of such a word will change according to the general tone-changing rules while the tone of the first character will change according to Table 11 [楊秀芳 , 1991].

Table 11. *Tone-Changing Rules for the First Character of Triple-Character Adjectives*

Original	1	2	3	4	5	7	8
Changed Tone	5	1	2	9(ptk)2(h)	5	5	*(ptk)5(h)

5.3 Evaluation

The same test data were used to evaluate the Tone Sandhi problem. To avoid other sources of error, we only considered correctly translated words. There were 5,576 such characters. The tones of 4,602 characters changed correctly. That is, the correct rate was 82.53%. After we examined the experimental results, we found the following types of errors:

(a) Within a phrase

Because the syntactic structures were not used, the tones of many nouns in noun phrases did not change while many verbs, in contrast, did change. An example is shown below:

(14) 台中_Nc 市地 _Na 重劃 _VC 弊案 _Na

(C) tai7 tiong7 chhi3 te7 tiong7 oe3 pe2 an3
(D) tai3 tiong chhi3 te7 tiong3 oe3 pe2 an3

(b) Object was a clause

The objects of some verbs were clauses. Since the clause behind such a verb was very long, its last tone did not change. For example,

(15) 再 _D 進而 _Cbb 要求_VF 提供 _VD 機密性 _Na 的 _DE 公務 _N 或 _Caa 國防 _Na 資料 _Na

(C) chia chin2 chit pou7 iau7 kiu5 the3 chhut9 ki7 bit seng3 e7 kong7 bu7 ia3 si3 kong kok9 hong5 e7 chu7 liau7
(D) iu3 ko2 chin2 li3 iau7 kiu3 the3 kiong7 ki7 bit siN3 e3 kong7 bu7 hek kok9 hong5 chu7 liau7

When a long clause was followed by the verb " 要求 ", it retained its last tone in the correct answer.

(c) **Nominalized verbs** like " 圖利 ", " 公辦 ", and " 重劃 " retained their last tones.

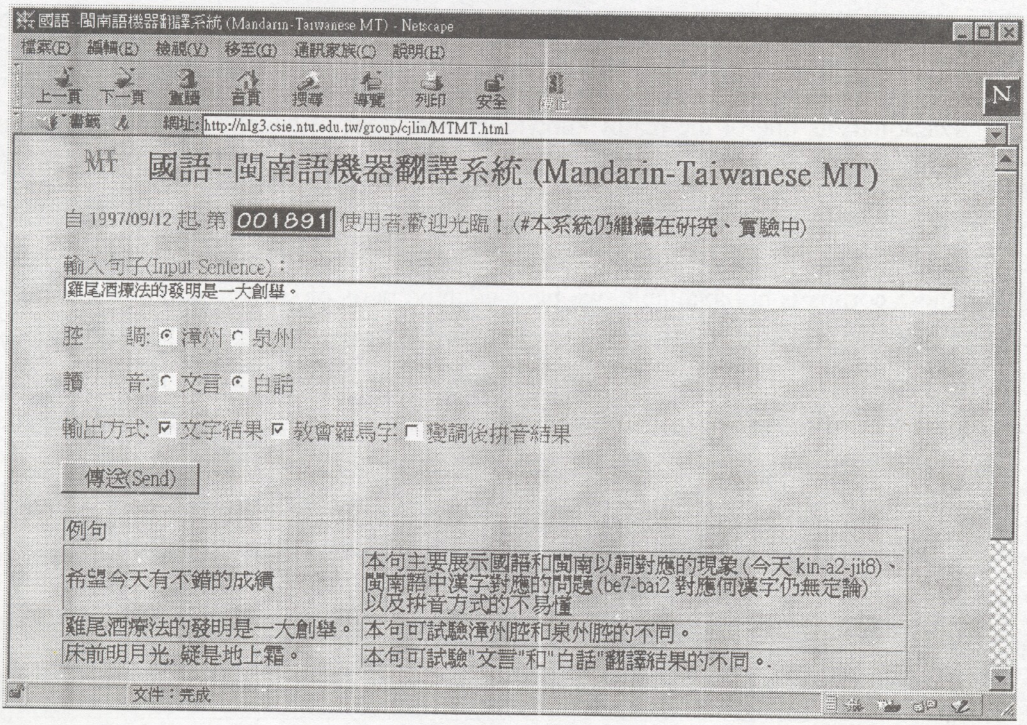
(d) **In predicate-object structure words** like " 修法 " and " 掃白 ", the second character was like an object, so it tended to retain its tone.

6. An Experimental System on the Web

An experimental system is demonstrated on the web site:

http://nlg3.csie.ntu.edu.tw/group/cjlin/MTMT.html

Figure 3 illustrates the home page of this system. After giving an input sentence, users can choose which accent is preferred. Besides, if the input sentence is selected from classical literatures or poems, it is preferable in the literary readings. The option " 讀音 " provides a selection " 文言 ". In the normal case, " 白話 " is chosen.


國語--閩南語機器翻譯系統 (Mandarin-Taiwanese MT) - Netscape
檔案(F) 編輯(E) 檢視(V) 移至(G) 通訊家族(C) 說明(H)
上一頁 下一頁 重讀 首頁 搜尋 導覽 列印 安全 停止
書籤 網址:
文件：完成

http://nlg3.csie.ntu.edu.tw/group/cjlin/MTMT.html

國語--閩南語機器翻譯系統 (Mandarin-Taiwanese MT)

自 1997/09/12 起, 第 001891 使用者,歡迎光臨！(#本系統仍繼續在研究、實驗中)

輸入句子(Input Sentence)：

雞尾酒療法的發明是一大創舉。

腔　　調: ● 漳州 ○ 泉州

讀　　音: ○ 文言 ● 白話

輸出方式: ☑ 文字結果 ☑ 教會羅馬字 □ 變調後拼音結果

傳送(Send)

例句	
希望今天有不錯的成績	本句主要展示國語和閩南以詞對應的現象(今天 kin-a2-jit8)、閩南語中漢字對應的問題(be7-bai2 對應何漢字仍無定論)以及拼音方式的不易懂
雞尾酒療法的發明是一大創舉。	本句可試驗漳州腔和泉州腔的不同。
床前明月光, 疑是地上霜。	本句可試驗"文言"和"白話"翻譯結果的不同。.

Figure 3 *The Home Page of the Mandarin-Taiwanese MT System*

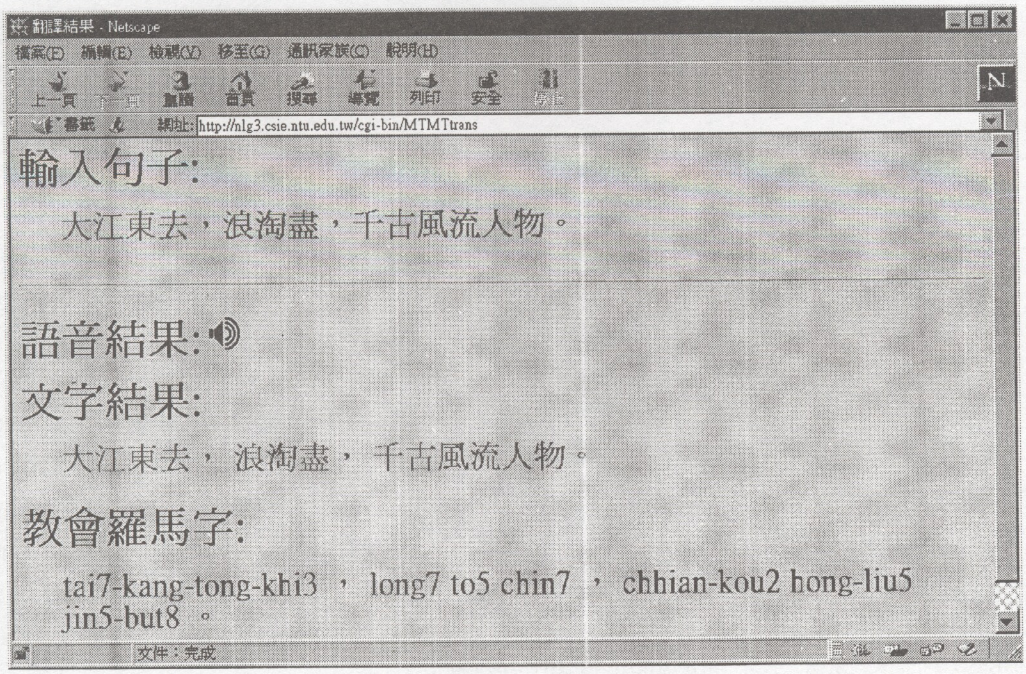
There are three kinds of ways to display the output results. That is, Chinese characters, Missionary Romanization and tone-changed form. Figure 4 gives an example, which is in classical literature reading. It is the first three sentences of the poem " 念奴嬌 " written by a famous Chinese poet 蘇軾 . We read it character by character. If this sentence is in colloquial reading, some characters will be translated into multi-character words. In that case, some will be translated wrongly. For example, " 大 " is read as "toa7" instead of "tai7". Both examples in Figures 5 and 6 are in colloquial readings but in different accents. Figure 5 shows the example of Chang-Chou accent and Figure 6 is in Chuan-Chou accent. The characters " 雞 " and " 尾 " have different readings, and the tones of " 療 " and " 的 " are changed differently.

7. Concluding Remarks

This paper has presented the first Mandarin-Taiwanese machine translation system in the world. We adopted an integrated approach in designing our MT system. In our system, a rule-based segmentation system divides a Chinese sentence into a sequence of tokens. A statistics-based tagger assigns a part-of-speech to each token. A lexical selection system chooses target lexical items. A speech synthesizer generates the speech output.

The major problems we have tackled are the unknown word problem, lexical disambiguation problem, and tone sandhi problem.

Dictionary is one of the major knowledge resources in developing such a system. The coverage of the Mandarin-Taiwanese dictionary and the inconsistencies in the tagging result are major errors in lexical selection. Setting up a large-scale dictionary will be indispensable for future research. Because a large-scale segmented and tagged Taiwanese corpus is not available, we have adopted parts of speech, and preferences in selecting suitable lexical items. The association of lexical items has not been considered in this paper. However, context information is important for making lexical choices. Further experiments are needed to integrate context knowledge. The tone sandhi phenomena in Taiwanese are complex. Although this study employed several rules for capturing such phenomena, there were still many errors. Knowledge of syntactic structure does help. How to incorporate the parsing results will be investigated further.


翻譯結果 - Netscape
檔案(F) 編輯(E) 檢視(V) 移至(G) 通訊家族(C) 說明(H)
上一頁 下一頁 重新載入 首頁 搜尋 導覽 列印 安全 停止
書籤 網址: http://nlg3.csie.ntu.edu.tw/cgi-bin/MTMTtrans
文件：完成

輸入句子:

大江東去，浪淘盡，千古風流人物。

語音結果:

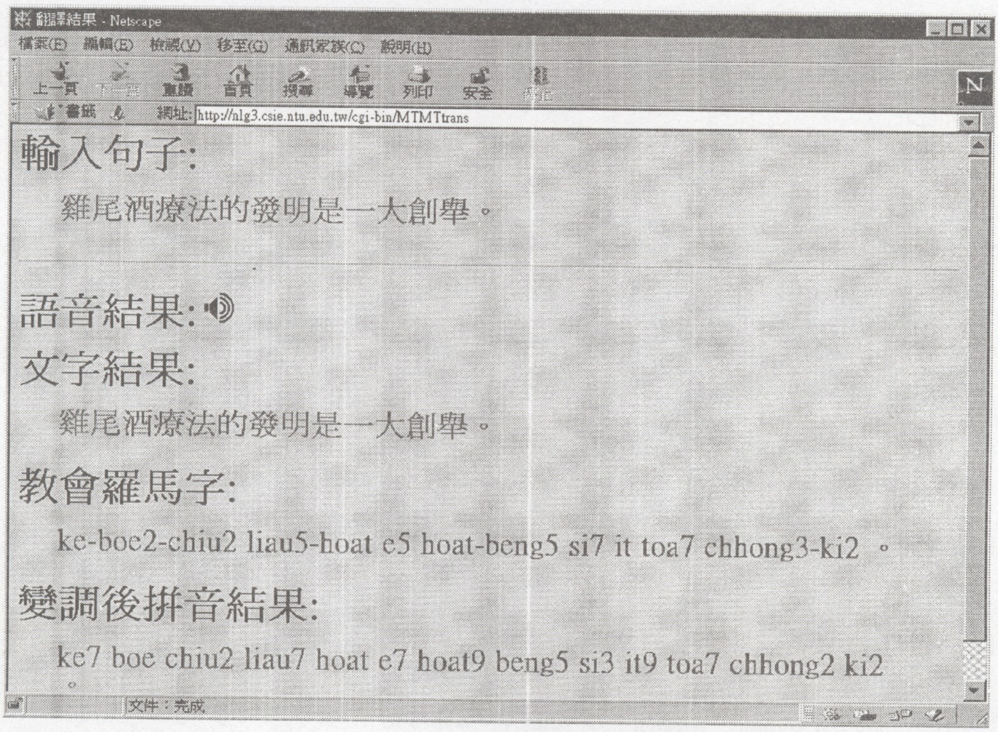
文字結果:

大江東去， 浪淘盡， 千古風流人物。

教會羅馬字:

tai7-kang-tong-khi3 ， long7 to5 chin7 ， chhian-kou2 hong-liu5 jin5-but8 。

Figure 4 *An Example in Classical Literature Reading*


翻譯結果 - Netscape
檔案(F) 編輯(E) 檢視(V) 移至(G) 通訊家族(C) 說明(H)
上一頁 下一頁 重讀 首頁 搜尋 導覽 列印 安全 停止
書籤
N
文件：完成

網址: http://nlg3.csie.ntu.edu.tw/cgi-bin/MTMTtrans

輸入句子:

雞尾酒療法的發明是一大創舉。

語音結果:

文字結果:

雞尾酒療法的發明是一大創舉。

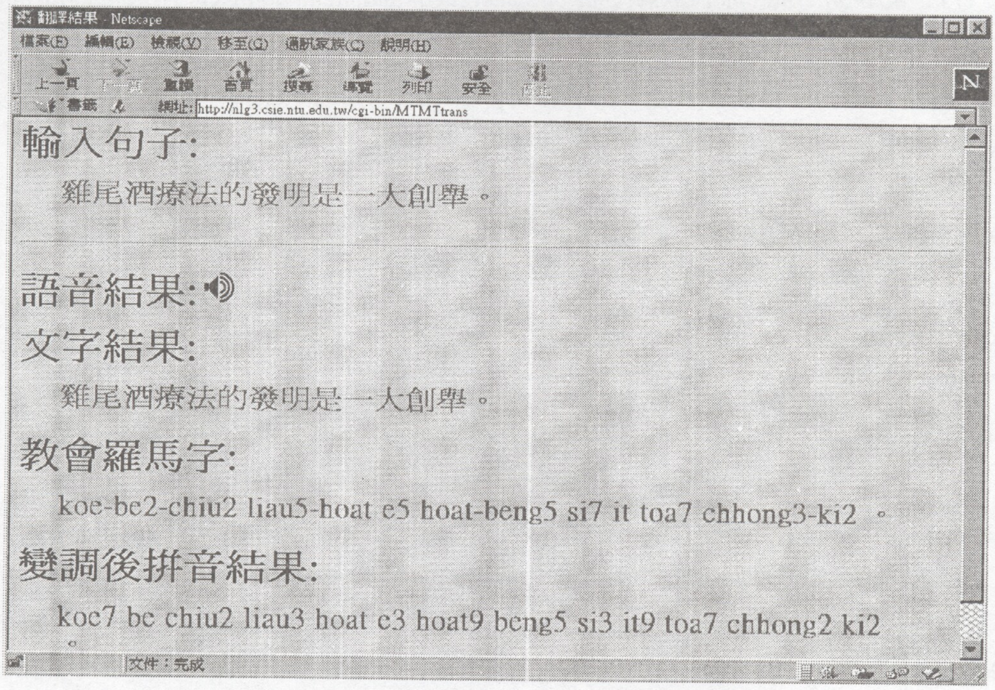
教會羅馬字:

ke-boe2-chiu2 liau5-hoat e5 hoat-beng5 si7 it toa7 chhong3-ki2 。

變調後拼音結果:

ke7 boe chiu2 liau7 hoat e7 hoat9 beng5 si3 it9 toa7 chhong2 ki2 。

Figure 5 *An Example in the Chang-Chou Accent*


翻譯結果 - Netscape
檔案(F) 編輯(E) 檢視(V) 移至(G) 通訊家族(C) 說明(H)
上一頁 下一頁 重讀 首頁 搜尋 導覽 列印 安全 停止
書籤
N
文件：完成

網址: http://nlg3.csie.ntu.edu.tw/cgi-bin/MTMTtrans

輸入句子:

雞尾酒療法的發明是一大創舉。

語音結果:

文字結果:

雞尾酒療法的發明是一大創舉。

教會羅馬字:

koe-be2-chiu2 liau5-hoat e5 hoat-beng5 si7 it toa7 chhong3-ki2 。

變調後拼音結果:

koe7 be chiu2 liau3 hoat e3 hoat9 beng5 si3 it9 toa7 chhong2 ki2 。

Figure 6 *An Example in the Chuan-Chou Accent*

Acknowledgements

We are very thankful to Professor Robert L. Cheng for providing us with the Taiwanese-Mandarin Dictionary.

References

Baker, K. *et al.* (1994), "Coping with Ambiguity in a Large-Scale Machine Translation System," *Proceedings of COLING-94*, Kyoto, Japan, 1994, pp. 90-94.

Bennett, W. and Slocum, J. (1985), "The LRC Machine Translation System," *Computational Linguistics,* Vol. 11, No. 2-3, 1985, pp. 111-119.

Brown, P. et al. (1990), "A Statistical Approach to Machine Translation," *Computational Linguistics*, Vol. 16, No. 2, 1990, pp. 79-85.

Chao, Y.R. (1968), *A Grammar of Spoken Chinese*, University of California Press, 1968, pp. 13-14.

Chen, H.H. and Lee, J.C. (1996), "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9, 1996, pp. 222-229.

Chen, K.H. and Chen, H.H. (1995), "Machine Translation: An Integrated Approach," *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, July 5-7, 1995, pp. 287-294.

Chen, K.H. and Chen, H.H. (1996), "A Hybrid Approach to Machine Translation System Design," *International Journal of Computational Linguistics & Chinese Language Processing*, 1(1), 1996, pp. 159-182.

CKIP (詞庫小組) (1995), " 研究院語料庫的內容及說明 ," *中文詞知識庫小組技術報告 #95-02*, 中央研究院 , 1995.

Grimes, B.F. (1996), *Ethnologue: Languages of the World*, Thirteenth Edition, Summer Institute of Linguistics, Inc., Dallas, Texas, 1996.

Lin, M.Y., Chiang, T.H., and Su, K.Y. (1993), "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING-93*, pp. 119-141, 1993.

Mitamura, T., Nyberg, E. and Carbonell, J. (1991), "An Efficient Interlingua Translation System for Multilingual Document Production," *Proceedings of Machine Translation Summit III*, Washington, DC, 1991.

Nagao, M. (1984), "A Framework of Mechanical Translation between Japanese and English by Analogy Principle," *Artificial and Human Intelligence* (Elithorn, A. Eds.), 1984, pp. 173-180.

中國大百科全書 (1988), 語言文字編，中國大百科全書出版社，北京，1988, pp. 227-228.

呂淑湘 (1980), *現代漢語八百詞*，商務印書館，北京，1980.

洪維仁 (1996), "臺灣閩南語輕聲變化與語法的關係," *《臺灣閩南語概論》講授資料彙編*，1996, pp. 330-382.

教育部 (1979), *常用國字標準字體表*，正中書局，台北，1979.

楊秀芳 (1991), *臺灣閩南語語法稿*，大安出版社，台北，1991.

楊青矗 (1992), *國台雙語辭典*，敦理出版社，台北，1992.

鄭良偉 (1989), *國語常用虛詞及其台語對應詞釋例*，文鶴出版有限公司，1989.

鄭良偉 (1993), *精速台語羅馬字練習與規律*，旺文社，台北，1993.

Appendix

Some of the results are included here to demonstrate the translation performance. Each set has one sentence in Mandarin, a corresponding Taiwanese sentence in the test data, and a translation output by our system (which received input from the word segmentation system and POS tagging system).

1. 希望 _VK 核四 _Nc 能夠 _D 延後 _VC 兩 _Neu 個 _Nf 月 _Na 再 _D 開工 _VH ， _VAC
hi-bong7 hek8-su3 ian5-au7 nng7 ko3 geh8 khai-kang
hi-bong7 hek-su3 e7-tit ian5-au7 nng7 ko3 geh8 chai3 khai-kang ，

2. 以便 _Cbb 朝野 _Na 之間 _Ng 能夠 _D 利用 _VC 這 _Nep 兩 _Neu 個 _Nf 月 _Na 的 _DE 時間 _Na 再 _D 評估 _VE 、 _VAC
i2-pian7 tiau5-ia2 e7-tang3 li7-iong7 che nng7 ko3 geh8 e5 si5-kan ko2-chai3 pheng5-kou2
i2-pian7 tiau5-ia2 chi-kan e7-tang3 li7-iong7 che nng7 ko3 geh8 tek si5-kan chai3 pheng5-kou2 、

3. 行政院 _Nc 將 _D 會 _D 立即 _D 和 _P 經濟部 _Nc 來 _D 研究 _VE 是否 _D 可以 _D 延後 _VC 核四 _Nc 開工 _VH 的 _DE 日期 _Na 。 _VAC
heng5-cheng3-iN7 chiong e7 sui5-si5 kah keng-che3-pou7 lai5 gian2-kiu3 si7-m7-si7 e7-tang3 ian5-au7 hek8-su3 khai-kang jit8-ki5
heng5-cheng3-oan7 chiong e7-hiau2 lip8-chit kah Keng-che3-pou7 lai5 gian2-kiu3 si7-m7-si7 e7-tang3 ian5-au7 hek-su3 khai-kang tek jit8-ki5

4. 台電 _Nc 將 _D 會 _D 損失 _VJ 大 _VH 筆 _Nf 的 _DE 賠償金 _Na 。_Na
tai5-tian7 chiong sun2-sit toa7 pit e5 poe5-siong2-kim
Tai5-tian7 chiong e7-hiau2 sun2-sit toa7-pit e5 pe5-siong5-kim 。

5. 雖然 _Cbb 申請 _VF 釋憲 _VA 的 _DE 立委 _Na 張俊雄 _Nb 等 _Cab 人 _Na 認為 _VE 總統 _Na 應該 _D 到庭 _VA 說明 _VE ,_Na
sui-jian5 sin-chhiaN2 sek-hian3 e5 lip8-ui2 tiuN-chun3-hiong5 in jin7-ui5 chong2-thong2 eng3-kai ai3 kau3-teng5 soat-beng5
sui-jian5 sin-chheng2 sek-hian3 e5 lip8-ui2 tiuN-chun3-hiong5 teng2 lang5 lin7-ui5 chong2-thong2 eng3-kai kau3-teng5 seh-beng5 ,

6. 但是 _Cbb 總統府 _Nc 一方 _Nh 仍然 _D 是 _SHI 以 _P 行政院 _Nc 法務部 _Nc 代表 _VK 出席 _VC 。_VAC
tan7-si7 chong2-thong2-hu2 iu5-goan5 si7 i2 heng5-cheng3-iN7 hoat-bu7-pou7 tai7-piau2 chhut-sek8
tan7-si7 Chong2-thong2-hu2 it hong iu5-goan5 si7-i2 heng5-cheng3-oan7 Hoat-bu7-pou7 tai7-piau2 chhut-sek8 。

7. 在 _P 法界 _Nc 和 _Caa 政界 _Nc 引發 _VC 了 _Di 重大 _VH 爭議 _Na 的 _DE 副總統 _Na 是否 _D 可以 _D 兼任 _VG 行政院長 _Na 的 _Na 問題 _Na ,_Na
ti7 hoat-kai3 kah cheng3-kai3 in2-hoat tiong7-tai7 cheng-gi7 e5 hu3-chong2-thong2 si7-m7-si7 e7-eng7=e kiam-jim7 heng5-cheng3-iN7-tiuN2 e5 bun7-toe5
chai7 hoat-kai3 kah cheng3-kai3 in2-hoat ah0 tiong7-tai7 cheng-gi7 e5 hu3-chong2-thong2 si7-m7-si7 e7-eng7=chit kiam-jim7 heng5-cheng3-oan7-tiong5 tek bun7-te5 ,

8. 由 _P 立委 _Na 張俊雄 _Nb 等 _Cab 人 _Na 申請 _VF 大法官 _Na 會議 _Na 解釋 _VE 後 _Ng ,_Na
iu5 lip8-ui2 tiuN-chun3-hiong5 chiah-e5 lang5 sin-chheng2 tai7-hoat-koaN hoe7-gi kai2-soe7 liau2-au7
iu5 lip8-ui2 tiuN-chun3-hiong5 teng2 lang5 sin-chheng2 toa7-hoat-koaN hoe7-gi7 kai2-sek au7 ,

9. 申請 _VF 釋憲 _VA 的 _DE 代表 _Na 馮定國 _Nb 提出 _VC 詢問 _VE ,_Na
sin-chheng2 sek-hian3 e5 tai7-piau2 pang5-teng7-kok the5-chhut sun5-bun7
sin-chheng2 sek-hian3 e5 tai7-piau2 Pang5 teng7 kok te5-chhut sun5-bun7 ,

10. 法務部長 _Na 廖正豪 _Nb 到底 _D 是 _SHI 代表 _VC 總統府 _Nc 還是 _Caa 代表 _VC 行政院 _Nc ,_Na
hoat-bu7-pou7-tiuN2 liau7-cheng3-ho5 tau3-toe2 si7 tai7-piau2 chong2-thong2-hu2 ia2-si7 tai7-piau2 heng5-cheng3-iN7
hoat-bu7-pou7-tiong5 liau7-cheng3-ho5 tau3-te2 si7 tai7-piau2 Chong2-thong2-hu2 hoan5-si7 tai7-piau2 heng5-cheng3-oan7 ,