Automatic Conversion from Phonetic to Textual Representation of Cantonese: The Case of Hong Kong Court Proceedings

Benjamin K. Tsou, K.K. Sin, Samuel W. K. Chan, Tom B. Y. Lai, Caesar Lun, K. T. Ko, Gary K. K. Chan, Lawrence Y. L. Cheung

Language Information Sciences Research Centre,
City University of Hong Kong,
Tat Chee Avenue, Kowloon,
Hong Kong SAR, China
Fax: (852) 2788-9734

Email: rlbtsou@uxmail.cityu.edu.hk.

ABSTRACT

The resumption of sovereignty over Hong Kong by China and the implementation of legal bilingualism there have given rise to an urgent need for producing verbatim court records of proceedings conducted in Cantonese, the predominant Chinese dialect spoken by the majority of the population. This has created a challenge to build up the jurilinguistic infrastructure vital for the full implementation of bilingualism and the retention of the Common Law system in Hong Kong. While there are Computer-Aided Transcription (CAT) systems for processing English and Mandarin (Putonghua), none exists for processing Cantonese. This paper discusses the design of a Cantonese CAT system based on the special features of Cantonese speech sounds. The CAT system works on the conventional English-based keyboard to process Cantonese and meets the bilingual requirements of the Hong Kong courts. By utilizing primarily statistical techniques, the system is highly successful in handling the ambiguity resolution of homophonous Chinese characters, a tantalizing problem in the conversion from phonetic to textual representation of Chinese. Additional linguistic analysis and related processing are discussed which could further improve the performance of the system from about 92% to over 94% accuracy.

1. INTRODUCTION

With the implementation of legal bilingualism in Hong Kong, Cantonese Chinese is increasingly used in the legal domain, particularly in court proceedings. Previously, when court proceedings were conducted exclusively in English, verbatim records were kept by court stenographers using the Pitman method and, more recently, the shorthand machine. The shorthand codes recorded by the machine were transcribed into English words via a Computer Aided Transcription (CAT) system. However, it is incapable of processing Cantonese, the dialect spoken by the predominant majority of Chinese litigants in Hong Kong. In the absence of an efficient and reliable device, the Judiciary of Hong Kong is confronted with the urgent problem of finding a way to maintain legally tenable records of court proceedings conducted in Cantonese Chinese. [1, 2, 3] Currently, the Judiciary has to resort to a primitive solution, i.e., transcribing the audio records of court proceedings into Chinese characters by means of audio-typing. This process is not only time-consuming but also error-prone.

To remedy the situation, two supporting facilities are required, namely, a computer-compatible Cantonese shorthand method that allows court stenographers to make verbatim records of Cantonese speech, and a Cantonese CAT system that facilitates the transcription of Cantonese shorthand codes into Chinese characters. Both English- and Mandarin-based CAT systems have been available for some years. [4, 5] However, neither a computer-compatible Cantonese shorthand method nor a Cantonese CAT is currently available. This paper discusses (1) a phonetically-based Cantonese shorthand method for use on stenograph machines; and (2) the design and initial implementation of a Cantonese CAT system capable of converting the phonetically-based shorthand codes into Chinese characters.

The rest of this paper is organized as follows. The next section briefly reviews the conversion from phonetic to textual representation in CAT. Section 3 gives a detailed account of the design of our Cantonese shorthand method and Cantonese CAT. Section 4 discusses the statistical techniques employed. Section 5 reports the evaluation results of the system. Section 6 discusses further linguistic analysis and enhancement features. Finally, Section 7 provides a summary and considers ways in which the system can be refined and expanded.

2. COMPUTER AIDED TRANSCRIPTION SYSTEM (CAT)

2.1 Overview of CAT

The main function of CAT is to transcribe shorthand codes into the words of a target language. Functionally, a CAT system can be divided into three major components: (a) a stenograph machine with a shorthand scheme, (b) an automatic transcription system (ATS), and (c) a supporting module for post-editing. (Figure 1)

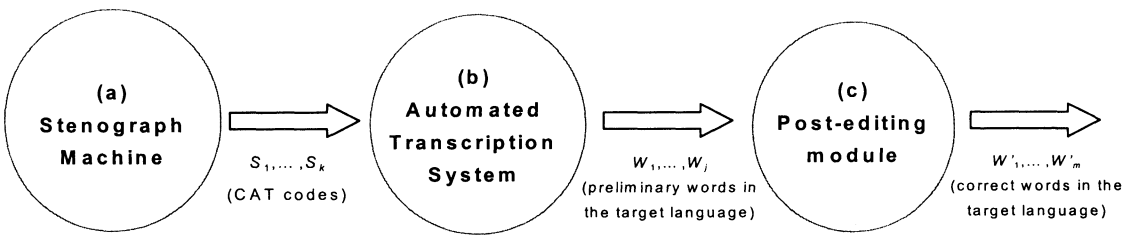

Figure 1: Schematic diagram of a typical CAT system.

The *stenograph machine* is a typewriter-like electronic device specifically designed for fast shorthand recording. To facilitate rapid key striking, it is built with a special key arrangement and has far fewer keys than the conventional typewriter keyboard. The machine is designed in such a way that several keys can be pressed simultaneously in one stroke. Thus there are many combinations of keystrokes within a single strike. The stenographer encodes speech as a sequence of *CAT codes*[1] simultaneous to the litigant speaking. The CAT codes are stored electronically for subsequent offline automatic transcription. The *automatic transcription system* (ATS) is a computer program that converts CAT codes into human-readable words of the target language (e.g. English words or Chinese characters). Finally, a CAT system normally comes with a supporting *post-editing module*. It enables stenographers to correct typos and mis-transcribed items and do further refinement to the output text.

Each CAT system comes with a shorthand method designed specifically for a particular language. This is referred to as *machine-compatible shorthand method*, in contrast to traditional hand-written shorthand schemes. Two existing machine-compatible shorthand systems were studied, namely, the Computer-Compatible Stenograph Theory (CCST) [4] for English and the Ya Wei Chinese Stenography (YWS) [5] for Mandarin. Both CCST and YWS are phonetically-based. The set of CAT codes represents the set of syllable inventory in the target languages. Each input scheme has its own keyboard. Basically CCST operates on a one-stroke-one-syllable basis whereas YWS works on a one-stroke-two-syllable basis.

Phonetically-based input inherently leads to ambiguous CAT codes. Each of these codes represents a

[1] In this paper we shall use "CAT codes" and "shorthand codes" interchangeably.

whole class of homonyms, words of identical pronunciation but different morphemes, resulting in code ambiguity. In both CCST and YWS, additional rules on using CAT codes have been enforced to eliminate code ambiguities. The rules in CCST generally appeal to the spelling of the intended English word, forcing the stenographer to supply extra "spelling hints" when typing a potentially ambiguous CAT code. In YWS, some 2,000 predefined special CAT codes are provided, each of which will enable the stenographer to uniquely identify the intended Chinese word. In short, both CCST and YWS reduce CAT code ambiguity through the provision of rules with which the stenographer can get unambiguous CAT codes.

2.2 The Homocode Problem – Preliminary Analysis

In a purely phonetically-based CAT code scheme, there is an isomorphic mapping from syllables of the target language to the CAT codes. However, because of the presence of homonyms in the target language, the mapping from phonetic representation to textual representation is one-to-many. We call this the *homocode problem* in the conversion from phonetic to textual representation. Figure 2 gives a more abstract view of CAT in view of this problem.

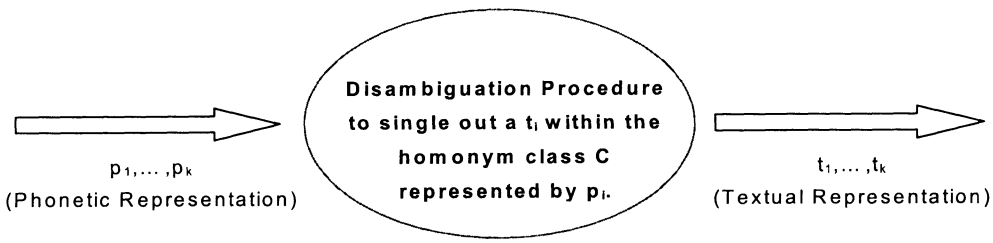

Figure 2: A more abstract view of CAT.

To circumvent the homocode problem, both CCST and YWS have not adopted a purely phonetically-based representation in the design of their CAT codes. As mentioned in the previous section, CCST provides heuristics for encoding extra information into their CAT codes to reduce ambiguity whereas YWS appeals to a large group of pre-defined codes to uniquely identify thousands of commonly used Chinese words.[2] Our present focus is whether methods of disambiguation similar to these can be adopted to overcome the homocode problem in Hong Kong's *Cantonese Chinese* setting.

Our observation is that it is difficult to adopt a scheme like that of YWS for use in Hong Kong's court environment. Even if a shorthand scheme similar to YWS could be devised, where Cantonese syllable types were represented on a one-keystroke-many-syllable basis on a YWS-like keyboard, the scheme would be very unintuitive for the court stenographers in Hong Kong because the stenographers are proficient in the English CCST tradition. Further, the significant linguistic gap between Mandarin and Cantonese also makes this not viable.[3] Neither can we adopt a simple CCST-like code conversion process by requiring the stenographer to key in "spelling hints" because Chinese characters are not based on spelling. A new synthesis is needed. It is noteworthy that the use of a purely phonetically-based shorthand scheme involves cognitive functions quite different from processing "spelling hints", or recalling ad hoc codes for a large list of salient items. In a purely phonetically-based input scheme, stenographers need to learn only how to phonetically represent syllables in CAT codes without worrying about ambiguity. Therefore, no ad hoc rules are needed. This makes Cantonese stenography far easier to learn and use, thus reducing the cost of training.

---

[2] To be more precise, both CCST's and YWS's heuristics try to achieve a one-to-one relationship between CAT codes and words in their target languages. Homonyms with identical pronunciations but different morphemes are now encoded with different CAT codes, making the relationship between CAT codes and syllables to become many-to-one.

[3] This gap can be as high as 40% in the lexical component. See [6].

3. CANTONESE CAT SYSTEM: SYSTEM DESIGN

This section describe in detail the design of our Cantonese CAT system, specifically addressing the homocode problem. We shall consider two core components: the Cantonese shorthand method and the ATS engine for the CAT system.

3.1 Keyboard Layout

Existing court stenographers have accumulated much experience in using CCST-based CAT system. It was decided that the CCST-based shorthand keyboard should be retained in view of two considerations. First, the stenographers are familiar with the CCST-based CAT system. It should be preserved as far as possible so as to capitalize on the existing equipment and the experience in the system. Second, English will continue to play a significant part in the legal domain. In the foreseeable future, the court proceedings are likely to be in a mixture of Cantonese and English. The Cantonese keyboard layout and the corresponding shorthand system must not just cater for Cantonese but should be capable of accommodating the bilingual transcription environment. Retaining the existing English CAT keyboard enables the stenographers to easily switch between the Cantonese and English environment, and minimizes the cost of such an extension.

3.2 Phonetic Representation of Cantonese—A Proposed Cantonese Shorthand Method

The keyboard layout being determined, the next step is to introduce a set of Cantonese phonetic symbols fully representable by key combinations on the CCST-based keyboard. The Cantonese romanization system *Jyutping* [8] has been chosen as the foundation for representing Cantonese syllable types in this regard. Jyutping has 19 initial consonants, 9 vowels and 8 codas, and is able to represent Cantonese speech sound with accuracy and consistency by phonetic symbols close to Pinyin and the International Phonetic Alphabet. Furthermore, all the phonetic symbols of Jyutping are representable on the CCST-based stenograph keyboard by a fairly natural extension. Instead of adapting a new keyboard and a set of novel key assignments, our Cantonese extension capitalizes on the existing CCST key combinations for English syllables. All original key combinations for English initial consonants, vowels and final consonants are preserved in the new scheme. Consonants and vowels unique to Cantonese are added to the scheme. The extension enables every Jyutping syllable type to be represented on a one-stroke-one-syllable basis, preserving the one-to-one correspondence between Jyutping syllables and CAT codes. The extended CCST scheme is called *CVC* (representing "***C***onsonant", "***V***owel" and "***C***oda" (i.e., the final consonant) of a Jyutping syllable).

3.3 Conversion from Phonetic to Textual Representation - The Problem of Homocodes

The adoption of the CCST-based keyboard and the CVC scheme requires that the associated ATS module is to be redesigned for the following reasons. Recall that in both CCST and YWS, the stenographer may invoke various disambiguation rules to obtain alternative unambiguous CAT codes. The design of YWS's disambiguation rules is tied to the Mandarin-oriented keyboard and differs fundamentally from the CCST-based one. This makes it infeasible to utilize the existing ATS to process CAT codes under the CVC scheme. Being an ideographic and basically monosyllabic language, Cantonese has many homophonous characters sharing identical syllable types (and therefore shorthand codes), and yet it is impossible to differentiate among these homonyms using spelling cues during online transcription because Chinese characters are not based on spelling. We call this the *Cantonese homocode problem* in the conversion from phonetic to textual representation of Cantonese. [4]

---

[4] The total inventory of Cantonese syllable types is about 720, and there are at least 14,000 Chinese character types. We estimated this for the legal domain with reference to a 0.85-million character corpus comprising mostly of court proceedings. 565 distinct syllable types were found, representing 2,922 distinct character types. Of the 565 syllable types, 470 have 2 or more homophonous characters. These 470 syllables represent 2,810 character types (which account for 94.7% of the corpus' tokens) each of which has at least 1 homonym. The homocode problem thus is indeed very serious in the domain of local court

We tackle the Cantonese homocode problem by equipping the Cantonese CAT system with statistical knowledge and an ambiguity resolution engine. In this way, the burden of homocode disambiguation is shifted from the stenographer (as in the cases of CCST- and YWS-based schemes) to the ATS. By adopting the bigram model commonly employed in speech recognition technology, the resolution engine is able to transcribe CAT codes into Chinese characters satisfactorily with over 90% accuracy. More precise evaluation figures will be presented in Section 5.

4. AMBIGUITY RESOLUTION OF HOMOPHONOUS CANTONESE CHARACTERS

We now turn to the statistical ambiguity resolution technique employed. Let us consider the homocode problem in more detail. The goal is to transcribe a sequence of shorthand codes $s_1, \ldots, s_i, \ldots, s_k$ into the intended Chinese character sequence $c_1, \ldots, c_i, \ldots, c_k$. In the phonetically-based shorthand system, a given shorthand code, s_i represents the Cantonese syllable of the intended Chinese character c_i. The homocode problem arises as the code s_i can be mapped onto any member of the homophonous character set $C = \{c_{i_1}, \ldots, c_{i_h}\}$, where $c_i \in C$ and every member of C shares the same shorthand code.[5] To resolve the ambiguity, we resort to statistical frequencies obtained from large training corpora, and search for the *most probable* Chinese character in context, for each shorthand code in the input code sequence. We seek to maximize the conditional probability in (1).

(1) $\quad P(c_1, \ldots, c_k \mid s_1, \ldots, s_k)$

where $\quad c_1, \ldots, c_k$ stands for a sequence of k Chinese characters, and

$\quad s_1, \ldots, s_k$ stands for a sequence of k input shorthand codes.

To compute (1) directly from corpus statistics, however, is impractical, as huge amount of data is required to generate any reasonable estimates of (1). For this reason, we look for a more practical approximation. We first rewrite (1) into (2) using Bayes' rule.

(2) $$\frac{P(c_1,\ldots,c_k) * P(s_1,\ldots,s_k \mid c_1,\ldots,c_k)}{P(s_1,\ldots,s_k)}$$

Now the goal is to find a Chinese character sequence $c_1,\ldots, c_k$ to maximize (2). The denominator $P(s_1, \ldots, s_k)$ can be ignored in the process, as its value remains the same for whatever character sequence chosen. The numerator in (2) can be approximated by (3) using two more assumptions[6].

(3) $\quad \prod_{i=1,\ldots,k} (P(c_i|c_{i-1}) * P(s_i|c_i))$

The maximal value of (3) can now be evaluated more practically. This is achieved by computing the co-occurrence frequencies obtained from a large training corpus (tagged with Cantonese syllable shorthand codes). The frequencies are to approximate the factors $P(s_i|c_i)$ (i.e., the pronunciation probability)[7] and also $P(c_i|c_{i-1})$ (i.e., the bigram probability). The approximated *maximal* value of (3) is efficiently computed using the Viterbi algorithm [9] to determine the best sequence of Chinese characters $c_1,\ldots, c_k$ as the output. This ambiguity resolution method, originally developed mainly for speech recognition [10, 11], has been found to be useful and has been built into our system's ATS module.

---

[5] There are 6 tone contours in Cantonese. However, tone contour is not encoded in transcription to reduce cognitive burden, thus making the homocode problem more acute.

[6] Both are "Markov assumptions" about historical influence. **Assumption 1: (Bigram model)** The bigram model assumes that for each Chinese character c_i in the target Chinese character sequence $c_1, \ldots, c_k$ we seek to obtain, the only historical factor of concern to its occurrence is the immediately previous character c_{i-1}. (When $i = 1$, this "historical factor" is stipulated as c_i's being the beginning of the discourse in question.) Accordingly the expression $P(c_1, \ldots, c_k)$ in (2) is approximated by $\prod_{i=1,\ldots,k} P(c_i|c_{i-1})$. **Assumption 2: (Independence of Pronunciation)** We assume that the way c_i is pronounced is independent of that for the preceding or succeeding members in $c_1, \ldots, c_k$. Accordingly the expression $P(s_1, \ldots, s_k|c_1,\ldots, c_k)$ in (2) is approximated by $\prod_{i=1,\ldots,k} P(s_i|c_i)$.

[7] The value of $P(s_i|c_i)$ need not always be 1 in Cantonese as c_i combines with different characters to form polysyllabic words.

5. EVALUATION

We have built several prototypes of the Cantonese CAT system for evaluation. The most basic version is the one equipped only with the basic ambiguity resolution method as described in section 4. More sophisticated prototypes were built upon this basic version by the addition of enhancement features. In this section, we will first describe the reference corpus used in the evaluation tests, followed by the description of the basic prototype and test results. Other enhanced prototypes will be discussed in section 6.

5.1 Corpora Used in Evaluation

We conducted experiments to evaluate the prototypes by setting up two data sets: *a training set* for training the ATS, and a *testing set* for evaluating the transcription accuracy. Both sets are derived from the corpus of authentic court proceedings (Chinese transcripts) obtained from the Hong Kong Judiciary.[8] Basic figures for the two sets of data are given in Table 1.

Data Set	Chinese characters	Numerals	Punctuation marks	English Words	Total
Training	715,501	8,786	99,287	20,491	844,065
Testing	175,735	1,359	22,662	5,606	205,362
Total	891,236	10,145	121,949	26,097	1,049,427

Table 1: The Training and Testing Data Sets

The Training Set

Recall that ATS requires pronunciation probability, $P(s_i|c_i)$, and bigram probability, $P(c_i|c_{i-1})$ in order to perform ambiguity resolution. To this end, we compiled a corpus of about 0.85 million Chinese characters, all tagged with the corresponding Jyutping syllables. The corpus was transformed into the training set by systematically assigning an appropriate CAT code for each such Jyutping syllables, and listing this code side by side with the original Chinese character. Supplied with this sequence of training data, the ATS estimated both the pronunciation and bigram probabilities for a given Chinese character, by computing the relative frequencies.[9]

The Testing Set

The testing set was used for simulating the stenographer's actual input on the stenograph machine in order to test the system's accuracy. The corpus consists of about 0.21 million Chinese characters. The testing set was obtained by replacing each Chinese character with the appropriate CAT code. The trained ATS took only the CAT code sequence as input and transcribed them into Chinese characters.

5.2 Basic Results

The same training and testing sets were used in conducting all the evaluation tests for every prototype.

---

[8] The case types of these court proceedings are heterogenous: they comprise *traffic*, *assault*, *robbery*, among others.

[9] Hence, based on this tagged corpus, an approximated value of $P(s_i|c_i)$ can be computed as the ratio of c_i's being pronounced as the syllable denoted by s_i, to the observed total occurrences of c_i'; similarly, $P(c_i|c_{i-1})$ is approximated as the ratio of observed occurrences of c_i after c_{i-1} to the observed total occurrences of c_{i-1}.

CAT_{VA}[10]

CAT_{VA} is the most basic prototype. It was subject to training with successively more inclusive subsets of the same training set, each containing the immediately previous one until the whole set is exhaustively used. To get the baseline reference, we also built a control, CAT_0, which converts a CAT code s_i into a Chinese character simply by selecting the member out of the homophonous set that has the highest occurrence frequency in the training set. The performance of the prototypes is summarized in Table 2.

Prototypes (training size in million characters)	CAT_0 (0.85)	CAT_{VA} (0.2)	CAT_{VA} (0.35)	CAT_{VA} (0.5)	CAT_{VA} (0.63)	CAT_{VA} (0.73)	CAT_{VA} (0.85)
Accuracy	78.0%	89.4%	91.2%	91.8%	92.1%	92.3%	**92.4%**

Table 2: Summary of CAT_{VA}'s performance.

As shown in Table 2, CAT_0 results in an accuracy of about **78.0%** whereas CAT_{VA} achieves at least **89.4%**, yielding over 11% increase in accuracy with as meagre a training subset as 0.2-million characters. With the use of full training set, CAT_{VA} reaches **92.4%** accuracy. At this level, there is a 14% gain in accuracy over CAT_0.

6. ADDITIONAL LINGUISTIC PROCESSING

The above discussion indicates that the best results that probabilistic information retrieval means could produce are unlikely to go substantially beyond 92% accuracy. Further efforts to enlarge the training corpus led to diminishing returns, as Table 2 indicates. Our subsequent investigation has shown that the accuracy can be improved more profitably by equipping the basic prototype with extra heuristic features on top of the statistical resolution engine. They include shallow linguistic processing to deeper linguistic analysis. These features are discussed below.

6.1 Generic Treatment of Numerals

In the original CCST-based stenograph keyboard, numeral tokens are input by the stenographer using the numeral row at the top of the keyboard. The shorthand codes for numeral types, e.g. 1998, 250,000, 2, 20, are thus represented by the numeral types themselves instead of being phonetically-based. This saves the stenographer from the tedious conversion of Arabic numerals into phonetically-based shorthand code during online input. However, it prevents the CAT system from capturing some potential regularities of numerals in the corpus. For example, the Chinese numeral is often followed by a member of a limited set of classifiers or units of measurement, e.g. sheet (for paper), dollar, metre. If each numeral type is represented by a different code, such regularities shared by *all* numerals can hardly be captured by the bigram probabilities between characters, as the frequency of individual numeral types is quite low even in a comparatively large corpus.[11] This is referred to as the *under-representation problem* of numerals. The negative impact of this problem during transcription is that numerals not encountered during training will *always* not be available in the statistical estimation and be treated as sparse data, affecting the accuracy. To overcome the problem, we represent all numeral types by using a generic category *NUM*. During the training phase, each time when a numeral token is encountered, *NUM*'s total frequency and the relevant probability figures will be updated. In this way, the shorthand method remains unchanged from the perspective of the stenographer, while the system's ATS module is able to track the bigram

[10] "VA" denotes the Viterbi algorithm.

[11] Numeral tokens constitute about 1% (about 10,400 tokens) of our 1 million-token training/testing corpus. Out of these tokens, however, only about 793 distinct numeral types are found in the same corpus. It is quite clear that even a much larger corpus won't lead to a substantially higher coverage rate of numeral types.

probabilities between this generic category and other Chinese characters during disambiguation. This *generic treatment of numbers* has been incorporated into our first prototype to yield an enhanced version, CAT_{VA+NUM}.

CAT_{VA+NUM}

We originally conjectured that by correcting the under-representation of numerals with the NUM category, CAT_{VA+NUM} could bring noticeable accuracy improvement. This has turned out not to be the case. We conducted the same evaluation tests on CAT_{VA+NUM}. The original transcription accuracy did not show improvement: with the full training set employed the accuracy still stayed around **92.4%**. The reason is that the majority of the numeral types in the test set have already been covered by the full training set, but the exact correlation between coverage and accuracy is currently still being worked out.[12] When more novel numerals not found in the training set are encountered in new test data, we believe CAT_{VA+NUM}'s performance will not be negatively affected, while that of CAT_{VA} will probably show degradation. We are still conducting experiments in this connection.

6.2 Further Analysis on Mis-transcription

Further detailed empirical analysis into possible cause of transcription errors revealed that in the CAT_{VA+NUM}'s evaluation, there were problematic Chinese characters that ranked high in terms of mis-transcription counts, and not just those that had the highest error ratios.[13] The statistics of mis-transcription is summarized in Table 3.

Size of Testing Data Set (characters)	Mis-transcribed Character Types (approx.)	Mis-transcribed Character Tokens (approx.)	Average Correctness (CAT_{VA+NUM})
205,362	1,500	15,000	92.4%

Table 3: Summary of Mis-transcription of CAT_{VA+NUM} (with the full training set of 0.85 million characters employed.)

As can be seen from Table 3, we need to narrow down the 1,500 distinct character types and identify the most problematic group. Our study revealed that, 33% of the 15,000 mis-transcription (about 2% of the whole testing set) was due to a group of about 40 characters, each with over 60 mis-transcription.[14] Moreover, there were two main reasons contributing to their errors: (1) High mis-transcription frequency solely attributable to high occurrence frequency of the character in question.[15] (2) High mis-transcription frequency due to the character's sharing of shorthand code with other homophonous characters with *much higher* occurrence frequencies, interfering with the correct selection during statistical disambiguation.[16]

---

[12] Among the 793 numeral types in our training/testing corpus (see footnote 11), 629 types are from the training set and 164 types from the testing set. We found altogether about 100 overlapping types, leaving fewer than 64 novel numeral types uncovered *in the testing set*. Curiously, these uncovered numeral types are usually of quite peculiar nature: such as the type "2001", and how adequately representative is the test corpus may be called into question.

[13] Pragmatically it would be useless to raise the correctness of a character *c* from 0% to 100%, if *c* had, say, only 1 occurrence within the 0.21 million character testing set.

[14] Mis-transcription counts range from 60 to about 700. The mean is 120 mis-transcription per member.

[15] An example of this is the Chinese character for the Cantonese morpheme *hai* ("to be"), with over 8,600 occurrences though only having about a 4.2% mis-transcription rate, i.e., about 360 mis-transcription.

[16] An example is the Cantonese morpheme *hai* ("at"), a homophone of the previous "hai" (previous footnote), which gave over 700 mis-transcription and yet having only about 1,600 total occurrences. The interesting fact is that this latter *hai* was found to be mis-translated as the former 44% of the times, accounting for *all* of those 700 mis-transcription.

$CAT_{VA+NUM+SE}$

It is evident that, if corrected, the 40-members group of characters can bring about at least 2% gain in the overall accuracy. The goal now is to reduce the influences of (1) and (2), minimizing both the transcription errors of high-frequencies characters and also their interference with the others. To this end, we selected a small set of characters from this group and assigned them with alternative unique shorthand codes. The selected characters have high frequencies of occurrence and were found to consistently produce most interference in transcription.[17] We derived the prototype $CAT_{VA+NUM+SE}$ ("SE" denotes special, unique encoding). Having been trained using the full training set with special encoding scheme, $CAT_{VA+NUM+SE}$ has gained over 2% in accuracy, resulting in **94.7%**. Further experimentation with the selection and size of this set is still in progress.

6.3 Other Linguistic Considerations

Our CAT system assumes that an ideal *speaker-hearer* world exists in court proceedings. This is not the case in reality. The false starts and incomplete sentences are not all consequential but they affect the corpus characteristics. Any attempts to effectively deal with these issues and to improve the results require more authentic records of proceedings.

There are also basic structural properties of language which need to be identified before any attempt to process them in the system. Good examples are incorporation of multiple linguistic elements and tone change. [12] describes how extralinguistic factors affect what apparently have been considered unpredictable for productive tone change in disyllabic nouns. This could account for some of the contributing factors on homonymy. Furthermore, examples such as the incorporation of aspect particle into the verb with a concomitant tone change could have consequential implications in the legal domains:

(1) joek-gwo nei lo^{33} di cin heoi

If you take the money to…

"If you take the money to (do something)…"

(2) joek-gwo nei lo^{33}-zo^{35} di cin heoi

If you take [comp.asp] the money to…

"If you have taken the money to (do something)…"

(3) joek-gwo nei lo^{35}-[ ] di cin heoi

If you take the money to…

"If you have taken the money to (do something)…"

(3) is a common contracted form of (2), where the completive aspect particle zo^{35} has been incorporated into the verb lo^{33} ("to take") which assumes the tone of the aspect particle. An acceptance of (1) admits no guilt but an acceptance of (2) or (3) is open to legal interpretation on the admission of guilt. The unambiguous characterization of linguistic situations such as the above is clearly important in its own right. To handle these issues, the transcription procedure must incorporate some kind of linguistic analyzer module on top of the statistically-driven engine. These issues await further investigation in the next phrase.

7. CONCLUSION

Both monolingual English- and Mandarin-based CAT systems have been available for some years. But our Cantonese CAT system is the first of its own kind to combine (1) a phonetically-based stenograph keyboard which was originally designed for Indo-European languages; and (2) a systematic romanization

[17] The set comprises about 20 members and is readily manageable by the court stenographers. Their separate unique short codes are modeled after their original phonetically-based codes to make them easier to remember by the stenographer.

scheme for the ideographic language, Cantonese. This combination is conducive to the full implementation of bilingualism in the Judiciary of Hong Kong. Most importantly, it helps implement a viable computer-aided Chinese stenograph system for the efficient keeping of legally tenable records of bilingual court proceedings. At the same time, the proposed system also allows stenographers to capitalize on their proficient skills in English stenography, thus facilitating a smoother and faster transition to a bilingual legal environment.

To realize this combination, a number of problems have to be solved. The most critical one is the homocode problem inherent in the conversion from phonetic to textual representation of Cantonese, and for that matter, of any tonal language. By applying statistical techniques commonly used in speech recognition and by employing extra heuristics, the system has, to a large extent, resolved the problem, achieving an accuracy rate of 94.7% in the overall transcription.

We are currently considering two ways to further improve the CAT system. The first is to build a trigram data model for the disambiguation process. The second is to try to incorporate grammatical knowledge in disambiguation. While the trigram model generally gives better results than the bigram one in other related areas of application, the setting up of its actual data model requires much bigger legal corpora, to cope with potential sparse data. We are currently trying to look for methods of reasonable approximation. The second way requires that we look for grammatical regularities beyond those surface constraints discernable solely from syllable-based CAT codes. A basic part-of-speech tagging may be necessary on the training data, a topic which we are also investigating.

ACKNOWLEDGMENT

Support for the research reported here is provided mainly through the Research Grants Council of Hong Kong under Grant CERG 9040326.

REFERENCES

[1] B. K. T'sou. 1993. "Some Issues on Law and Language in the Hong Kong Special Administrative Region (HKSAR) of China." Language, Law and Equality: Proceedings of the 3rd International Conference of the International Academy of Language Law (IALL). (eds.) K. Prinsloo et al. Pretoria: University of South Africa. pp. 314—331.

[2] K. K. Sin and B. K. T'sou. 1994. "Hong Kong Courtroom Language: Some Issues on Linguistics and Language Technology." Paper presented at the Third International Conference on Chinese Linguistics. Hong Kong.

[3] S. Lun, K. K. Sin, B. K. T'sou, and T. A. Cheng. 1995. "Diannao Fuzhu Yueyu Suji Fangan." [The Cantonese Shorthand System for Computer-Aided Transcription] (in Chinese) Proceedings of the 5th International Conference on Cantonese and Other Yue Dialects. B. H. Zhan (ed). Guangzhou: Jinan University Press. pp. 217—227.

[4] M. Glassbrenner and G A. Sonntag. 1986. Computer-Compatible Stenograph Theory. 2 vols. Illinois:Stenogrph Corporation.

[5] Y. W. Tang. (ed.) 1985. Suji Jishu [Shorthand Technology] (in Chinese) Beijing: Xinhua Bookstore.

[6] Y. W. Tang. 1994. Yawei Zhongwen Suluji—Peixun Jiacheng. [Yawei Chinese Stenograph Machine —A Training Course] (in Chinese) Shehui Kexue Wenxian Chubanshe [Social Science Literature Publisher].

[7] B. K. T'sou. 1997. " 'Sanyan' 'Liangyu' Shuo Xianggang" ['Three Types of Speech' and 'Two Languages' in Hong Kong] (in Chinese) Journal of Chinese Linguistics, vol. 25, 1997, pp 290-307.

[8] Linguistic Society of Hong Kong. 1997. Yueyu Pinyin Zibiao [Cantonese Jyutping Transliteration Word List]. (in Chinese) Hong Kong : Linguistic Society of Hong Kong.

[9] A. J. Viterbi. 1967. "Error Bounds for Convolution Codes and an Asymptotically Optimal

Decoding Algorithm", IEEE Transaction on Information Theory 13: pp. 260—269.

[10] L R. Rabiner. 1989. "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proceedings of IEEE. Reprinted in A. Waibel and K F Lee. (eds.) Readings in Speech Recognition. San Mateo, CA: Morgan Kaufmann.

[11] L R. Rabiner and B. H. Juang. 1993. Fundamentals of Speech Recognition. Englewood Cliffs, N.J., PTR Prentice Hall.

[12] B. K. T'sou. 1994. "A Note on Cantonese Tone Sandhi (CTS) as a Diffusional Phenomenon", Festschrift for Prof. William S.Y. Wang on his 60th Birthday. (Ed) M. Chen and O. Tzeng, Pyramid Press, Taipei, pp.539—549.

[13] R. Bauer. 1988. "Written Cantonese of Hong Kong." Cashiers de Linguistique Asie Orientale 17-2: 245-293.

[14] P. Chan. 1993. Report of the Working Party on the Use of Chinese in Courts. Judiciary, Hong Kong.

[15] S. W. K. Chan and B. K. T'sou. (in press). "Semantic inference for anaphora resolution: Toward a framework in machine translation. Machine Translation".

[16] E. Charniak. 1993. Statistical Language Learning Cambridge MA, MIT Press.

[17] Y. F. Cheung. 1994. "A Study of Court Reporting in the Judiciary of Hong Kong." MA Diss. City Polytechnic of Hong Kong.

[18] Y. Guo. 1991. Xian Dai Shi Yong Su Ji (Contemporary Practical Shorthand). Changsha: Hunan Technology Publishing Co.

[19] K. K. Luk and O. T. Nancarrow. 1990. "Polyglossia in the `Print Cantonese' Mass Media in Hong Kong." Journal of Asian Pacific Communication 1-1: 27-43.

[20] L R. Rabiner. 1989. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of IEEE. Reprinted in Waibel and Lee (1990).

[21] L R. Rabiner and B. H. Juang. 1993. Fundamentals of Speech Recognition. Englewood Cliffs, N.J., PTR Prentice Hall.

[22] K. K. Sin. 1988. "Meaning, Translation and Bilingual Legislation." Proceedings of First International Conference on Language and Law. Ed. Paul Pupier and Jose Woehrling. Montreal: Wilson & LafleurLtee. pp. 509-515.

[23] K. K. Sin. 1992. "The Translatability of Law." Research on Chinese Linguistics in Hong Kong. Ed. Thomas H. T. Lee. Hong Kong: Linguistic Society of Hong Kong. pp. 86—100.

[24] K. K. Sin and B. K. T'sou. 1996. "Some Reflections on the Development of Language Rights in Hong Kong." Paper presented to the International Conference on Language Rights, University of Illinois, USA, April 1996.

[25] D. B. Snow. 1991. "Written Cantonese and the Culture of Hong Kong: the Growth of a Dialect literature." Diss. Indiana university.

[26] D. B. Snow. 1993. "Chinese Dialect as Written Language: The Cases of Taiwanese and Cantonese." Journal of Asian Pacific Communication 4-1: 15-30.

[27] Stenograph Corporation. 1993. Premier Power User Guide (2nd Version). Illinois: Stenograph Corporation.

[28] W. C. Suen. 1993. "Reflections on the Translation of Laws into Chinese and Bilingual Drafting in Hong Kong." Seminar on Bilingual Legislation in Hong Kong. University of Hong Kong.

[29] B. K. T'sou. 1994a. "Language Planning Issues Raised by English in Hong Kong: Pre-and Post-1997." English & Language Planning: A Southeast Asian Contribution. (ed.) T. Kandiah and J. Kwan-Terry. Singapore: Times Academic Press. pp. 192—217.

[30] B. K. T'sou. 1994b. "A Note on Cantonese Tone Sandhi (CTS) as a Diffusional Phenomenon", Festschrift for Prof. William S.Y. Wang on his 60th Birthday. Pyramid Press, Taipei, (Edited by M. Chen, O. Tzeng), 1994, pp.539-549.

[31] B .K. T'sou, T. B. Y. Lai, S. W. K. Chan, K. K. Sin, K. T. Ko, L. Y. L. Cheung and G. K. K. Chan. 2000. "Statistically-based Model for Computer-Aided Transcription Application" Paper accepted for presentation in the 5th International Conference on Statistical Analysis of Textual Data (Lausanne, Switzerland).

[32] B.K. T'sou, H. L. Lin and T. B. Y. Lai. 1997. "Human Judgement as a Basis for Evaluation of Discourse-Connective-based Full-text Abstraction in Chinese" in Proceedings ROCLING X Computational Linguistic Conference, Taipei, August 1997, pp.195-205.

[33] B. K. T'sou and K. K. Sin. 1992. "Hanyu, Cantonese and Hong Kong's Legal Language." Proceedings of the International Conference on Language Development in Macau During a Transition Period, Macau. pp. 164—170.

[34] B. K. T'sou, K. K. Sin, C. Lun and T. B. Y. Lai. 1990. "The Implications of Mass Media Language for Legal Language." Paper presented at the Annual Research Forum of Linguistic Society of Hong Kong.

[35] B. K. T'sou, K. K. Sin, C. Lun and T. B. Y. Lai. 1992. "Hong Kong's Future Legal Language: A Projection Based on a Study of Hong Kong's Media Language." Proceedings of the Third International Conference on the Teaching of Chinese. Taipei., pp. 29—48.