TRANSCRIBING SOUTHERN MIN SPEECH CORPORA WITH A WEB-BASED LANGUAGE LEARNING SYSTEM

Jun Cai[1,2], Jacques Feldmar[1], Yves Laprie[1], Jean-Paul Haton[1]

[1] Groupe Parole, LORIA-CNRS & INRIA, BP 239, 54600 Vandoeuvre-les-Nancy, France
[2] Dept. of Cognitive Science, Xiamen Univ., 361005 Xiamen, China

ABSTRACT

The paper proposes a human-computation-based scheme for transcribing speech corpora. The core idea of the scheme is to implement a Web-based language learning system to collect orthographic and phonetic labels from a large amount of language learners and use some criteria to choose the commonly input labels as the transcriptions of the corpora. It is essentially a technology of distributed knowledge acquisition. The benefit of the scheme is that it makes the transcribing task neither tedious nor costly. The design of a system for transcribing Min Nan speech corpora is described in detail.

Index Terms— Speech transcription, southern Min (Min Nan) language, distributed knowledge acquisition, Web-based language learning

1. INTRODUCTION

Southern Min (a.k.a. Min Nan or Southern Fujian, or "闽南话" in Chinese,) language refers to a family of Chinese dialects which are spoken mainly in southern Fujian, eastern and southwestern Guangdong—both of which are coastal provinces in Mainland China—and neighboring areas. The geographic distribution of Min Nan also includes Taiwan and some areas in Southeast Asia. It is usually called Taiwanese by residents of Taiwan. In 2005, the total number of its speakers is estimated at 49 millions. In common parlance, Min Nan usually refers to Xiamen dialect (better known as the Amoy language) because Xiamen (Amoy) is the principal city of southern Fujian and Amoy accent is considered the most important, or even the standard accent in all variants of Min Nan. The Amoy dialect has played an influential role in history, especially in the relations of Western nations with China, and was one of the most frequently learned of all Chinese languages/dialects by Western people during the second half of the 19th century and the early 20th century.

Although Min Nan is a widely used language in southeastern China, up to now no practical large vocabulary speech recognition system for it has been developed in Mainland China where the research and development efforts in speech recognition have been focusing on applications in Mandarin. In Taiwan, researches on speech recognition for Min Nan began at the end of 1990's and some large vocabulary Min Nan speech recognition systems had been successfully developed since then [1]. As the first attempt in Mainland China to develop a Min Nan LVCSR system, the Speech Group at Xiamen University had collected a set of recordings of radio news broadcast in Min Nan for 150 hours and started on building acoustic models for Min Nan speech recognition since Nov. 2007.

Transcribing the Min Nan speech recordings into both orthographic and phonetic forms is a resource- and labor-intensive procedure. Because transcribing speech corpora is intrinsically a task of adding annotations to a set of pieces of information and is similar to the task of labeling images, some basic ideas can be drawn from the human-computation-based games for labeling images [2] to deal with the difficulties in transcribing the Min Nan speech recordings. In this paper, we propose combining the task of transcribing Min Nan speech with the pedagogical procedure of Min Nan language learning, via designing a Web system based on the concept of Human Computation. Other than the image-labeling games from which the players' benefit is only for fun, our Web system provides a platform for Min Nan learners to facilitate their training in listening comprehension and pronunciation.

2. DIFFICULTIES IN TRANSCRIBING SPEECH CORPORA

In the context of speech recognition, two levels of transcriptions, namely orthographic transcription and phonetic transcription are indispensable for the initial training of acoustic models and language models, respectively [3, 4]. These two levels of transcriptions are, from the point of view of machine learning, data sets of human linguistic knowledge on which inductive learning is performed by the acoustic and language models. The generation of these transcriptions is essentially a knowledge acquisition procedure during which human linguistic knowledge is extracted from a certain source (human annotators or an existing ASR system, for example) and stored

in the corpus. Since the performance of a Min Nan speech recognition system relies heavily on the availability of a substantial amount of carefully and accurately transcribed Min Nan speech corpus, it is of great importance that the transcriptions are created in a high quality.

There are basically two alternative schemes for adding orthographic and phonetic transcriptions to a speech corpus: annotating manually or automatically. Manual annotation is performed by human annotators who have been trained in linguistics. This is a way to extract linguistic knowledge directly form human experts. Usually, all transcriptions should be cross-checked and verified by a group of annotators to correct any annotation mistakes. That means manual annotation is not only a tedious and time-consuming procedure, but also a costly project for the developers. Lacking available financial resource is the reason why manual annotation is mostly performed only on a small corpus (e.g. TIMIT) or on a small part of a large corpus. In our project, speech recordings for only 20 hours have been manually transcribed by a group of Min-Nan-speaking students.

Various automatic methods [4-7] have been proposed to add orthographic transcriptions and phonetic transcriptions to speech corpora and many of them can be used to generate transcriptions for Min Nan speech data. Automatic speech recognition (ASR) systems are usually applied to generate transcriptions. Although for carefully read radio news broadcasts, state-of-the-art ASR systems can achieve a word accuracy of more than 90% and a phone accuracy of more than 80%, this kind of application to orthographically and phonetically transcribe speech data is still far from being successful.

Also lexicon lookup methods can be used to generate phonetic transcriptions. These methods map orthographic transcriptions to their pronunciations based on a pronunciation dictionary. Lexicon lookup is not always effective for annotating Min Nan speech. A serious problem is that a pronunciation dictionary usually can not represent all possible different pronunciations and dialectal variations of a certain Min Nan word. In Min Nan, there is a phenomenon called "being pronounced differently in literary speaking and vernacular speaking" ("文白异读"). A character can be pronounced in a classical way when it is used literarily, while it can be pronounced in a totally different way when it is used in a vernacular speaking. Besides that, some characters have different pronunciations in different contexts. For example, the character "成" is pronounced as [sêng] in the word "成功", as [siâⁿ] in the word "几成", as [chiâⁿ] in the word "成做", and even as [chhiâⁿ] in the word "成家". Also, there is another problem that many proper names which happen in news are usually not included in the pronunciation dictionary. Due to these reasons, today's technology can not ensure that the quality of automatic generated phonetic transcriptions based on lexicon lookup is high.

3. HUMAN COMPUTATION AND ITS APPLICATION FOR LABELING IMAGES

Human Computation (or Human-based Computation) is a technique when a computational process performs its function via outsourcing certain steps to humans [8]. In traditional computation, a human provides a formalized problem description to a computer, then he receives a solution to interpret. In Human Computation (HC), however, the computer asks a person or a large number of people to solve a problem, then collects, interprets, and integrates their solutions. The basic idea of HC is that there are a lot of problems that humans can easily solve but computer can not yet. Some of these problems can be solved by just making good use of human processing power [2, 9]. HC outsources many operations of a typical computational algorithm to humans. As a result of this outsourcing, HC can process the representations for which there is no computational innovation operators available, for example, natural language.

The technique of HC has been successfully adopted to design interactive systems either to use human intelligence to perform tasks which appear to be difficult for computer programs to solve or to collect commonsense knowledge from the general public over the Web [2, 8-11]. Such interactive systems can be designed as computer games which people play for fun. A typical example is the ESP game which addresses the problem of image labeling [2]. The ESP game combines people's desire to be entertained together with the acquisition of meaningful labels for images.

The ESP game is designed to be played by two randomly paired partners and usually played online by a large number of pairs simultaneously. Players can not know who their partners are, nor are they allowed to communicate with each other. For players, the goal of the game is to guess what the partner is typing for each image. Once both partners have input the same textual string for a certain image while the image is on their screens (this situation is described as they "agree on the image"), the game moves on to the next image. Partners strive to agree on as many images in a fixed time period as they play the game.

By this way, the system can collect a set of strings for every image.

Since the players can not communicate with each other, the easiest way for both partners to input the same string is to type something related to the content of the image. The agreement by a pair of independent players implies that the label is probably meaningful. From the perspective of the system, the textual string on which two players agree is typically a good label for the image. Therefore, by using only the words that players agree on the system can ensure the quality of the labels. Furthermore, if a string has been agree by a lot of pairs of players, the probability that it could be a meaningful label would be high. So, a "good label threshold" (e.g., 40) can be used to further guarantee the quality of the labels.

The ESP game is much like an algorithm in the input/output behavior. Its input is a set of images, while the output is a set of labels that properly describe the images. This kind of game is usually referred to as the "game with a purpose" (GWP). Actually, such a game runs a distributed computation in people's brains instead of in silicon processors.

4. DESIGN OF THE SYSTEM

By drawing the basic idea of GWP, an HC-based Web system has been designed to deal with the problem of Min Nan speech transcription. A criticism of GWP is that it is unfair to the players since they spend a lot of time to perform some task but the benefit for them is only for fun. Our speech transcribing system combines the task of transcribing Min Nan speech with the pedagogical procedure of Min Nan language learning. So the users can use the system to learn Min Nan pronunciation while performing the task of transcribing the speech recordings. Note that pronouncing and understanding Min Nan syllables are especially difficult to Western learners, mainly because seven different tones can be attached to each syllable. A well-designed language learning system will be helpful to the learners.

The framework of the system is depicted in Fig. 1. The system consists of several modules among which the core part is a module for Web-based Min Nan language learning. It is this module that utilizes the technique of HC to realize the speech transcription. The input to the whole system is a set of speech sentences, while the output is the confident transcriptions of them. The system is described as follows.

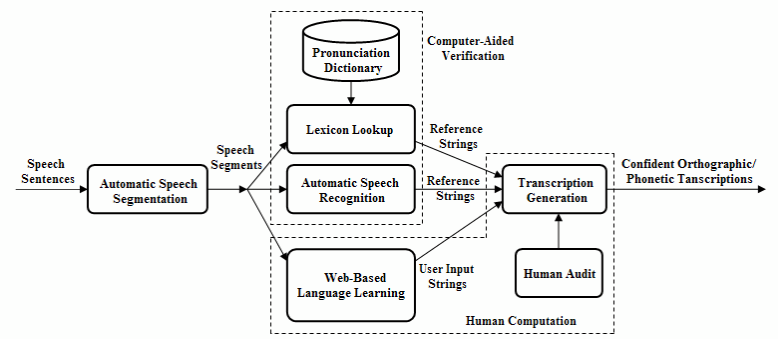

Fig. 1 Framework of the Human-Computation-based Transcription System

4.1. Automatic Segmentation

All the speech recordings are pre-partitioned into sentences of around 10s, i.e., the input speech data is stored as a set of speech sentences. A sentence of 10s is too long to transcribe, especially for beginning learners of Min Nan, In order to make the task less difficult, an automatic segmentation module has been developed to partition the input speech sentences into short segments of about 3s. In Min Nan news broadcast, a 1s piece of speech contains 4 valid syllables, on average. Like Chinese Mandarin, every syllable of Min Nan consists either of one syllable onset plus one syllable rime or of only one syllable rime, and every syllable corresponds to one and only one Chinese character. Therefore, the orthographic transcription of a speech segment of 3s contains 12 Chinese characters on average, and the corresponding phonetic transcription normally contains about 20 phonetic symbols. Labeling such segments would not be a difficult task for the users.

Since the SNR of the recorded news broadcast is high, we use short-time-energy-based voice activity detection technique [12] to perform the automatic segmentation. Short time energy is computed for every speech frame, and its value is utilized to

discriminate between speech and silence. The end point of a segment can be decided if there is a continuous silence of 100ms. Same as in Mandarin, the pronunciation unit in Min Nan is syllable, and there is normally no stop between the onset and the rime within a syllable. Continuous silence only happens between syllables. So, the end points decided with the above method normally locate between the utterances of different Chinese characters. That implies that there is no incomplete syllable in every speech segment generated by the automatic segmentation module. Every segment can be fully transcribed by a string of Chinese characters.

4.2. User Interface

The user interface of the Web-based language learning module provides a platform for the users to learn Min Nan language. The interface can play speech segments to users and enable them to practice listening comprehension and phonetic training. In addition, this module collects up the user input labels and stores the labels in a set of XML files (that will be described later). Based on a large amount of input labels, orthographic and phonetic transcriptions are generated for every speech segment by utilizing human computation mechanism.

Fig. 2 depicts the Web page (http://59.77.21.117:8080/humanComputation/jsp/minnan.jsp) for Min Nan phonetic training. The users' task here is to listen to speech segments and to input corresponding phonetic symbols. There is an audio player for playing and re-playing each segment. To the right of the audio player is a textbox which can display the corresponding text (if it has already been stored in the XML file) of the current speech segment in order to give the user some clues to understand the speech.

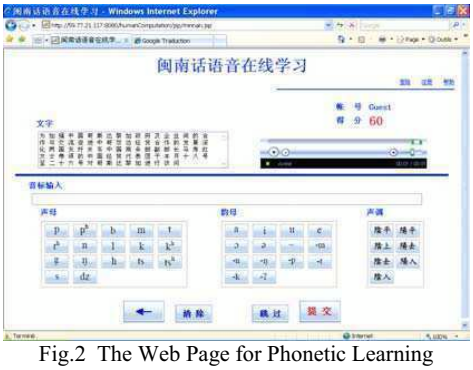

Fig.2 The Web Page for Phonetic Learning

Under the audio player and the textbox, there is an input box to enable the user to input phonetic symbols. In this system, we adopt "Romanization of Taiwan Min Nan Language Phonetic Alphabet" [13] to label the phonemes of Min Nan. This phonetic system uses Latin alphabet to mark the phonemes and it is close to the IPA symbol system. There are 17 syllable onsets, 14 syllable rimes, and 7 tones in the phonetic system. Special keyboard on the Web page is designed for inputting the phonetic marks. That means that the only way to input phonetic symbols is to use a mouse to click on the special keyboard. This design can also restrict the users' input in the set of the phonetic alphabet; no illegal symbol could be input as phonetic marks. By double-clicking a key, the system can play the standard pronunciation of its corresponding phoneme to help users master the pronunciation of each phoneme.

4.3. Transcription Storage

In the system implementation, the speech data is stored as sentences. To each speech sentence is attached an XML file to store transcriptions and other information about the sentence. Since each sentence is partitioned into segments and what the users transcribe is exactly segments of sentences, the information in the XML file is organized according to the segmentation of the sentence. The XML schema is shown below.

```
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION>
    <UTTERANCE> speech.file </UTTERANCE>
    <LENGTH> number of seconds </LENGTH>
    <TEXT> orthographic transcription of the sentence </TEXT>
    <SAMPLINGRATE> 16 <SAMPLINGRATE/>
    <WORDLENGTH> 16 </WORDLENGTH>
    <ENDIANNESS> little endian </ENDIANNESS>
    <NUMBER_SEGMENTS> n </NUMBER_SEGMENTS>
    <SEGMENT> segment01
      <FILENAME> segment.file </FILENAME>
      <START_TIME> start point </START_TIME>
      <SEG_LENGTH> number of seconds </SEG_LENGTH>
      <LABEL>
        <WORD_LABEL> word level annotation
            </WORD_LABEL>
        <PHONE_LABEL> phone level annotation
            </PHONE_LABEL>
      </LABEL>
      <ANNODATA> annotation01
        <WORD_LABEL> word transcription </WORD_LABEL>
        <WORD_CONFIDENCE> m </WORD_CONFIDENCE>
        <PHONE_LABEL> phonetic transcription
            </PHONE_LABEL>
        <PHONE_CONFIDENCE> m </PHONE_CONFIDENCE>
      </ANNODATA>
            …
      <ANNODATA> annotation20
        <WORD_LABEL> word transcription </WORD_LABEL>
        <WORD_CONFIDENCE> m </WORD_CONFIDENCE>
        <PHONE_LABEL> phonetic transcription
            </PHONE_LABEL>
        <PHONE_CONFIDENCE> m </PHONE_CONFIDENCE>
      </ANNODATA>
    </SEGMENT>
        …
        …
    <SEGMENT> segment10
      <FILENAME> segment.file </FILENAME>
      <START_TIME> time of start point </START_TIME>
```

```
        <SEG_LENGTH> number of seconds </SEG_LENGTH>
          …
    </SEGMENT>
</ANNOTATION>
```

4.4. Computer-Aided Verification

The "games with a purpose" (such as ESP and CYC's FACTory) use the number of players who agree on a piece of input string to decide the quality of the string. Though this mechanism is useful to collect common sense facts and knowledge, the Web-based game itself can not ensure the collected information or knowledge is absolutely correct. Sometimes, a common sense is likely to be a common error. In transcribing Min Nan speech, for example, if an incorrect or inaccurate pronunciation has been taught to a group of students, the phonetic marks of a speech sentence transcribed by these students are probably consistent with each other, but they are totally wrong. To prevent the system from outputting totally wrong transcriptions, we have introduced an automatic speech recognition (ASR) module and a lexicon lookup module to facilitate computer-aided verification of the input transcriptions.

The ASR module is built on the transcribed 20-hour subset of the speech data. The word level correctness and phone level correctness of the recognition of this module are about 80% and 64%, respectively. Every speech segment which is played to the user for being transcribed is also fed into the ASR module for being recognized automatically. For each speech segment, the module outputs a Chinese character string and a phoneme string as the recognition results. The system uses these two strings as references to verify the quality of the user input strings of characters and phonemes: after a label string is input by a user, it is compared with its reference counterpart. We use the Maximum Substring Matching algorithm [14] to compute the word error rate of the user input string, and define the consistency between the input string and its reference as follows:

$$Consistency = 1 - word\ error\ rate \quad (1)$$

If the user's input is a correct transcription string, the consistency between it and the reference string may well be high since the reference string is almost correct in the statistical sense; otherwise, if the user inputs a totally wrong string, the consistency is low. Therefore, the consistency of an input string can be used as a measure of its quality. In the current implementation, input strings with the consistency less than 40% are refused by the system. Thus a low-quality transcription has no chance to appear in the results of human computation, even though the transcription is a common sense.

The role of the lexicon lookup module is substantially the same as that of the ASR module, except that it only generates phonetic reference strings for the verification of users' input of phonetic strings. For each speech segment which orthographic transcription has been available, the module looks up the pronunciation dictionary to find the phoneme string corresponding to the orthographic transcription. If a unique phoneme string is found, it will be used as the reference of the input phonetic string. The consistency between these two strings is computed and then is used to evaluate the quality of the input phonetic string.

4.5. Collecting up Transcriptions

By collecting up a large amount of user inputs, the system utilizes HC techniques to generate orthographic and phonetic transcriptions of all speech segments. To describe the generation of transcriptions, we still take the procedure of phonetic training as the example. After the user logins the system, a sentence is selected randomly for the speech corpus. If the orthographic transcription of the sentence exits in the XML file, it will be displayed in the textbox to help the user understand the speech. Then, the system plays a segment of the sentence to the user and waits for the user's response. If the user has input a phoneme string in the input box and has pressed the "Done" button, the system is triggered to handle the input string.

In the processing, the confidence measure associated with each transcription string plays an important role to decide the number of users who agree on the transcription. For each input string, when it is stored in the XML file for the first time, its confidence measure is initialized to its consistency value computed with Eq. (1). Afterwards, every time the same string is input by a different user, its confidence measure is increased by 1. Therefore, the greater the number of users who agree on a transcription, the greater its confidence measure will be. Once every speech segment in the corpus has been repeatedly transcribed by a large amount of users, the best transcription can be decided based on the idea of HC principle: the transcription string with the maximum confidence measure is chosen as the best transcription of every speech segment.

However, the confidence measure chiefly acts as the indicator of the popularity of each transcription string. It does not necessarily ensure the quality of the transcription. If there are common errors in an input string and the consistency value of the string happens to be greater then the threshold (40%), the string can successfully enter the XML file. After being labeled by a large number of users, this transcription string

will stand high among all transcriptions of the same segment because the users commonly input it and it has the maximum confidence measure. So, this string with common errors will be chosen as the best transcription of the segment. The HC technique itself can not get rid of the situations that common errors survive in the final transcription. To prevent such situations, a human-audit module has been introduced into the system to facilitate human experts to selectively inspect the transcriptions. Bad transcriptions will be deleted in the XML files. Or, if a certain user makes mistakes frequently in transcribing, the human-audit module can be used to drive out all the transcriptions input by the specific user. Combining human audit by experts with human computation by a large mass of people, we can ensure the high quality of the final transcriptions of the speech corpus.

5. SUMMARY AND FUTURE WORK

This paper has proposed an HC-based Web application system which is utilized to generate orthographic and phonetic transcriptions of Min Nan speech corpora. The system combines speech data transcription with language learning. It adopts human computation to collect transcriptions from learners of Min Nan language, and uses human audit by Min Nan language experts to further guarantee the quality of the transcriptions. The experimentation of a prototype version of the system shows that the HC-based transcribing scheme is an effective and economical way to collect orthographic and phonetic labels. We believe that if the system is used by a large amount of people, high-quality transcriptions can be generated based on the collected inputs.

Several prospective research issue might be pursued in order to further improve the performance and the utility of the system, as well as to extend the application of this HC-based scheme for transcribing speech corpora. Firstly, more language learning functions should be added into the system to make it more helpful to the learners. Secondly, the incentive mechanism in computer games can be drawn into the transcribing system to attract more learners to be involved in the application and to motivate them to use the system more. We should also research on how to extend the application of this HC-based labeling scheme to transcribe English or French speech corpora. Some techniques, especially the segmentation methods in the current system are not valid for English or French speech processing. A prospective way to partition long speech sentences in English or French into short segments is that even this problem is outsourced to language learners and we segment the speech by using human computation.

6. REFERENCES

[1] R-Y. Lyu, Y-J. Chiang, and W-P. Hsieh, "A large-Vocabulary Taiwanese (MIN-NAN) Multi-syllabic Word Recognition System Based Upon Right-context-dependent Phones with State Clustering by Acoustic Decision Tree," *Proc. of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, Nov. 30–Dec. 4, 1998.

[2] L. von Ahn, and L. Dabbish, "Labeling Images with a Computer Game," *Proc. of the ACM SIGCHI conference on Human factors in computing systems*, Vienna, Austria, pp. 319–326, 2004.

[3] S. Bird, and M. Liberman, "A Formal Framework for Linguistic Annotation," *Speech Communication*, Vol. 33, Issue 1-2, Special issue on speech annotation and corpus tools, pp. 23–60, Jan. 2001.

[4] K. Demuynck, T. Laureys, and S. Gillis, "Automatic Generation of Phonetic Transcriptions for Large Speech Corpora," *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, pp. 333–336, Sept. 2002.

[5] S. S. Chen, E. Eide, and M. J. F. Gales, et al, "Automatic transcription of Broadcast News," *Speech Communication*, Vol. 37, Issues 1-2, pp. 69–87, May 2002.

[6] H. Y. Chan, P. Woodland, "Improving Broadcast News Transcription by Lightly Supervised Discriminative Training," *Proc. of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, Vol. 1, pp. 737–740, May 2004.

[7] S. Chang, L. Shastri, and S. Greenberg, "Automatic Phonetic Transcription of Spontaneous Speech (American English)," *Proc. of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, Vol. 4, pp. 330–333, Oct. 2000.

[8] A. Kosorukoff, "Human-based Genetic Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 5, pp. 3464–3469, Oct. 2001.

[9] P. Singh, T. Lin, and E. T. Mueller, et al, "Open Mind Common Sense: Knowledge Acquisition from the General Public," *Lecture Notes in Computer Science*, Vol. 2519/2002, pp. 1223–1237, Feb. 2004.

[10] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: A Game for Collecting Common-Sense Facts," *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, pp. 75–78, 2006.

[11] Amazon.com, Inc., "Amazon Mechanical Turk," *http://www.mturk.com*, 2005-2007.

[12] E. Dong, G. Liu, and Y. Zhou，et al, "Voice Activity Detection Based on Short-time Energy And Noise Spectrum Adaptation," *Proc. of the 6th International Conference on Signal Processing*, Vol. 1, pp. 464 – 467, Aug. 2002.

[13] Ministry of Education (Republic of China), "Romanization of Taiwan Min Nan Language Phonetic Alphabet (臺灣閩南語羅馬拼音方案)," *http://www.ntcu.edu.tw/tailo*, Oct. 2006.

[14] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory Algorithm and System Development*, Prentice Hall PTR, 1st edition, pp. 421, Apr. 2001.