

# Multiple Document Summarization Using Principal Component Analysis Incorporating Semantic Vector Space Model

Om Vikas\*, Akhil K Meshram\*, Girraj Meena\*, and Amit Gupta\*

## Abstract

Text Summarization is very effective in relevant assessment tasks. The Multiple Document Summarizer presents a novel approach to select sentences from documents according to several heuristic features. Summaries are generated modeling the set of documents as Semantic Vector Space Model (SVSM) and applying Principal Component Analysis (PCA) to extract topic features. Pure Statistical VSM assumes terms to be independent of each other and may result in inconsistent results. Vector space is enhanced semantically by modifying the weight of the word vector governed by Appearance and Disappearance (Action class) words. The knowledge base for Action words is maintained by classifying the words as Appearance or Disappearance with the help of Wordnet. The weights of the action words are modified in accordance with the Object list prepared by the collection of nouns corresponding to the action words. Summary thus generated provides more informative content as semantics of natural language has been taken into consideration.

**Keywords:** Principal Component Analysis (PCA), Semantic Vector Space Model (SVSM), Summarization, Topic Feature, Wordnet

## 1. Introduction

With the advent of the information revolution, electronic documents are becoming a principal media of business and academic information. The Internet is being populated with hundreds of thousands of electronic documents each day. In order to fully utilize these on-line documents effectively, it is crucial to be able to extract the main idea of these documents. Having a Text Summarization system would thus be immensely useful in serving this need. Multiple Document Summarization System aids to provide the summary of a document set that

---

\*Indian Institute of Information Technology and Management, Gwalior, India- 474010

E-mail: {omvikas, akhil, girrajmeena, amitgupta}@iiitm.ac.in

contains documents which belong to same topic. It can also be used to generate the summary of a single document.

In the present work, we propose a method of text summarization that uses semantics of data in order to form efficient and relevant summary. Summary is generated by constructing Statistical Vector Space Model (3.1) and then modifying it using the concept of Action words to form Semantic Vector Space Model (3.2). Action Words are identified using the Action Word Classifier which makes use of Wordnet [Kedar *et al.*] in order to analyze the semantics of word.

Principal Component Analysis (3.3) is then applied on SVSM to reduce the dimension of multidimensional data sets. Singular Value Decomposition (SVD) is carried out on SVSM as a part of PCA to yield singular values and eigen vectors. Backprojection is then performed to project the documents onto the eigen space yielding projected values of documents which are henceforth compared with the singular values to yield the most relevant document/topic. Sentence Extraction (3.4) from multiple document sets has been assigned weight on the basis of keywords obtained from the most important document/topic. Sentences with higher weight are taken to form a summary.

## 2. Related Work

Various multiple document summarization systems already exist. This document summarizer is based on Kupeic 95 [Kupeic *et al.* 1995] which is a method of training a Bayesian classifier to recognize sentences that should belong in a summary. The classifier estimates the probability that a sentence belongs in a summary given a vector of features that are computed over the sentence. It identifies a set of features that correspond to the absence/presence of certain words or phrases and avoids the problem of having to analyze sentence structure. Their work focused on analyzing a single document at a time. Since then, there has been lot of work on the related problem of Multiple-document Summarization [Regina *et al.* 1999; Radev *et al.* 1998], where a system summarizes multiple documents on the same topic. For example, a system might summarize multiple news accounts of the recent massacre in Nepal; into a single document. Our hypothesis is that the similarities and differences between documents of the same type (*e.g.* bios of CS professors, earnings releases, etc.) provide information about the features that make a summary informative. The intuition is that the 'information content' of a document can be measured by the relationship between the document and a corpus of related documents. To be an informative summary, an abstract has to capture as much of the 'information content' as possible. To gain a handle on the problem of capturing the relationship between a document and a corpus, we examined several papers on Multiple-Document Summarization [Regina *et al.* 1999; Radev *et al.* 1998, 2000, 2004; Otterbacher *et al.* 2002]. However, we found most of their approaches were not applicable to

our problem since they are mostly trying to match sentences of the same meaning to align multiple documents. The MEAD summarizer [Radev *et al.* 2000, 2001], which was developed at the University of Michigan and at the Johns Hopkins University 2001 Summer Workshop on Automatic Summarization, produces summaries of one or more source articles (or a ‘cluster’ of topically related articles).

Our Summarizer works on the documents belonging to same topic. It is strongly motivated by the analogy between this problem and the problem of face identification, where a system learns features for facial identification by applying PCA to find the characteristic eigenfaces [Turk *et al.* 1991; Pentland *et al.* 1994; Moon *et al.* 2001].

### **3. New Methodology**

Any set of documents dealing with the same subject is decomposed using Vector Space model. The important keywords can be extracted from the Vector Space Model using a threshold. Such keywords are called thematic keywords which are based on statistics. Important sentences can be extracted and a summary can be made using thematic keywords. We propose a new methodology for multiple document summarization by enhancing the VSM using semantics and identifying topic features based keywords to make the multiple document summary. The approach is:

1. Statistical VSM construction from Multiple Document Set.
2. Semantic VSM generation using the concept of Contextual Action Words using Wordnet.
3. Application of PCA on Semantic VSM to reduce the dimension of the multidimensional data set yielding the most important Keywords.
4. Score Sentences based on several features such as sentence length cut-off feature, position feature, keyword weight, etc.
5. Generation of Summary extracting the sentences with high Score.

#### **3.1 Statistical VSM Construction**

The Multiple Document Summarizer models the set of documents related to the same topic as the Statistical Vector Space Model based on several heuristics. The simplest way to transform a document into a vector is to define each unique word as a feature. The weight of a feature being decided based on the contribution of various parameters such as Cue-phrase Keywords, topic keywords and term frequency in document. The weight of the feature is being termed as Feature Combination.

The vector representations of the documents; collectively define an n-dimensional vector space (where each document is an nx1 vector). The m document vectors taken as the columns

of an  $n \times m$  matrix  $D$ , define a linear transformation into the vector space.

### 3.2 Semantic VSM Construction

The existing vector space model is statistical in nature. This vector space is input to a number of tools and processes like a summarizer and information retrieval system. PCA/SVD Technique has been applied earlier for Summarization based on statistical vector space [Gong *et al.* 2001]. Some times this statistically generated model is unable to define the context. Keywords identified by a statistical model can be non-contextual in nature. Therefore, an effort is to be made in the direction of identification of contextual keywords and modification of existing model so that it can be more helpful and contextual for various applications like text summarization and text retrieval.

To identify the contextual keywords, we try to exploit human psychology. In any article, we identify that those words are important which either give a sense of either appearance or disappearance of any object/event. Thus, after we have the pure statistical vector space we need to enhance the vector space semantically by modifying the weights of the word vector by identifying the Appearance and Disappearance (ACTION class) words. To do so, we need to have a knowledgebase (KB) with some seed wordlist which belongs to appearance or disappearance. Following, are the steps involved in the semantic vector space model.

1. Get the tf matrix,  $T$  from existing document  $D$ .
2. Identify the set of action words,  $A$  from the given tf matrix,  $T$ , (number of action words  $=n$ ).
3. Find the associated object list  $O_i$  for action word  $A_i$ ;  $A_i \in A$ ,  $0 < i < n$ .
4. Find contextual objects  $C_o$  from Object list  $O_1, O_2, \dots, O_n$ .
5. Modify weight of contextual objects in  $T$  to form semantic vector space  $S\{T\}$ .

#### 3.2.1 Identification of Action words

Action words are the backbone of the semantic vector space model.

*Definition: Action words are verbs that are used to strengthen the way experiences are presented whether it is expressing positive or negative experience.*

With the help of Wordnet, the terms from the tf (term frequency) matrix which belong to the ACTION class can be easily classified. The algorithm uses a seed word list to identify the action words.

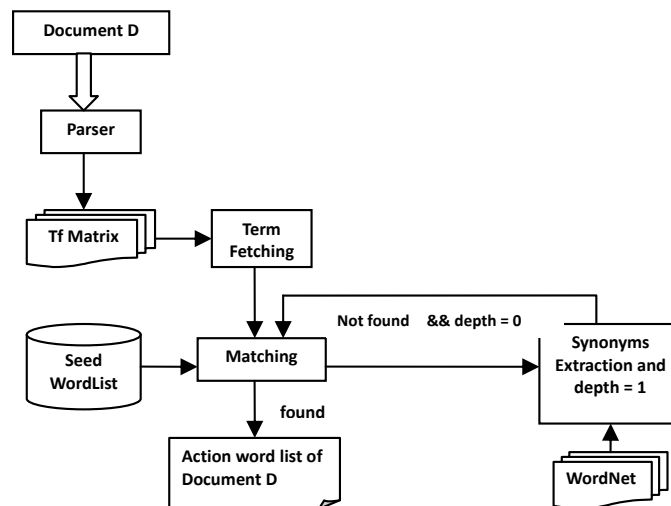
*Definition: Seed Word List is the collection of action words. (Appendix A)*

Whenever a term from the tf matrix is fetched, it is matched against seed word list. If it is matched, then the fetched term is action word; otherwise, synonyms of the fetched terms are matched.

```

Given Input: T = {t1,t2,...,tn}.
List type: A = { }, integer type: depth
Do: for every t ∈ T
    depth = 0;
    match(t,seedwrldst)
    if found then A = A U t.
    else, not found
        if(depth == 0)
            match(extractsynonym (t),seedwordlist)
            depth = 1
        else
            continue
    endif
endif
endfor
Output : A = {t1,t2,...tm}.

```



**Figure 1. ACTION Word Classifier**

To decide whether the word belongs to action list or not, we have to build a seed wordlist and compare them with standard meaning. For example, let ‘**devastation**’ be the word to be decided as action or not. After searching in WordNet, the following meanings were obtained:

- **desolation** (an event that results in total destruction)

- **ravaging**, (plundering with excessive damage and destruction)
- **destruction**, (the termination of something by causing so much damage to it that it cannot be repaired or no longer exists)

From the first and last meaning, it clearly lies in the phenomenon of appear/disappear so it will be appended into the seed list along with its Synset.

### 3.2.2 Finding the Objects of the Action

Merely acquiring the ACTION words doesn't provide the semantic to the vector space. We have to find whether these words are really important. The importance of the word can be estimated by the application of the word in the article. Objects corresponding to the Action Words and their weight in Statistical VSM have to be identified in order to determine the extent of relevancy of Action Words. The Objects are the Nouns or Adjectives for the Action. The nearest Noun for Verb is identified using POS Tagger and termed as Object of Action. Only those sentences are to be chosen which contain action words.

Today *broke* fire in Delhi. (Action is verb)

Today/NN *broke*/VBD **fire**/NN in/IN Delhi/NNP

Destruction of material *happens* due to this fire.

*Destruction*//NN of/IN **material**/NN happens/VBZ due/JJ to/TO this/DT fire/NN ./.

Many suffered from the *broken* glass in the road. (Action is Adjective)

Many/JJ suffered/VBD from/IN the/DT *broken*/JJ **glass**/NN in/IN the/DT road/NN. ./.

The authority *arrives* here soon.

The/DT **authority**/NN arrives/VBZ here/RB soon/RB. ./.

**Table 1. Action-Object List**

Action word	Objects
broke	fire, glass
arrives	Authority
destruction	material

The bold ones are selected as objects for the Action. The action-object list is prepared by the help of POS tagger and Contextual Action Words are determined.

### 3.2.3 Classification of Contextual Words

Contextual words are being defined as those action words which are applied to the important object. The Weight of Action Word is being taken as the maximum weight amongst all the objects corresponding to the given Action Word. The weight obtained is added to the weight of the corresponding Action Words in Statistical VSM yielding Semantic Vector Space Model for the given set of Documents.

If we take an example of a single document:

*Today broke fire in Delhi. Mass Destruction of material happens due to this fire. Many suffered from the broken glass in the road. The authority arrives here soon. Till now there is report of any casualties in these fire except from few injures. Thanks for the local communities for help.*

the Vector Space generated on the basis of term frequency feature is

**Table 2. Vector Space of Single Document (1 x 20)**

Broke	0.1889	Injuries	0.1889
Fire	0.5669	Thanks	0.1889
Delhi	0.1889	Local	0.1889
Mass	0.1889	communities	0.1889
Destruction	0.1889	Help	0.1889
Material	0.1889	Authority	0.1889
Happens	0.1889	Arrives	0.1889
Suffered	0.1889	Report	0.1889
Broken	0.1889	Casualties	0.1889
glass	0.1889	Road	0.1889

The Action Word List obtained corresponding to the above example is: broke, destruction, and arrives. Now, the Action-Object list is prepared by identifying the Object words in which the ACTION words are acted.

**Table 3. Object-Action List for example above**

Action word	Objects
broke	today, fire, glass
destruction	Material
arrives	Authority

Each ACTION word has been given weight as per the contextual word obtained corresponding to it.

**Table 4. Contextual Action List**

Action word	weight factor (wt)
broke	Max (0.1889,0.5669) = 0.5669
destruction	0.1889
arrives	0.1889

The Statistical VSM is now modified and the Semantic VSM is being generated as follows

**Table 5. Semantic Vector Space Model**

<b>Broke</b>	<b>0.7558</b>	Injuries	0.1889
Fire	0.5669	Thanks	0.1889
Delhi	0.1889	Local	0.1889
Mass	0.1889	communities	0.1889
<b>Destruction</b>	<b>0.3778</b>	Help	0.1889
Material	0.1889	Authority	0.1889
Happens	0.1889	<b>Arrives</b>	<b>0.3778</b>
Suffered	0.1889	Report	0.1889
Broken	0.1889	Casualties	0.1889
glass	0.1889	Road	0.1889

Similarly, the model is extended for multiple documents. This Semantic Vector Space Model is used further to determine important Keywords and henceforth, the summary.

### 3.3 Principal Component Analysis

Principal Component Analysis (PCA) [Michael *et al.* 2003] is used to reduce the multidimensional datasets to lower dimensions for analysis. Singular Value Decomposition (SVD) [Michael *et al.* 2003] is carried out on Semantic VSM to find the principal components of Vector Space. The singular value decomposition (SVD) of matrix  $A_{m \times n}$  is the factorization  $A=U\Sigma V^T$ , where  $U$  and  $V$  are orthogonal, and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ ,  $r = \min(m, n)$ , with  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$ . The columns of  $V$  are the 'hidden' dimensions that we are looking for. The diagonal of  $\Sigma$  are the singular values which are the weights for the new set of basis vectors.  $\Sigma$  is symmetric, its singular values are its eigen values and its basis vectors are the eigen vectors.

Given an eigen vector  $e$ , we can find the corresponding dimension in document space.

$$\vec{d} = D \cdot \vec{e}$$



After determining out the dimension of eigen vector in document space, backprojection of  $d^*$  is carried out. Commonly, composing a vector in terms of the principal components is called backprojection. Since our principal components or eigen documents are all orthogonal vectors, this is easy to accomplish. Let  $E$  be the matrix formed from the eigen documents then vector  $p$  is the document projected onto the eigenspace.

$$\vec{p} = E^T \vec{d}$$

Relevance of the topic/document is calculated by dividing projected component by the corresponding Singular Value. Metrics thus obtained is arranged in decreasing order excluding out the negative metrics. Main topic/Document is the one with highest metric value.

After selecting the main topic, we now need the topic keywords. We simply take the eigen document vector corresponding to main document and select the words with high weight. These are the set of Keywords which are of high relevance in summary.

### 3.4 Sentence Extraction

To identify sentences that should belong to summary, several features have been taken into consideration.

- **Sentence-Length Cut off Feature** – If the sentence length is greater than 4 words, only then it is taken into consideration.
- **Position Feature** – Sentences have been given some weight based on their position in the paragraph whether it is in initial, middle or final.
- **Keywords** - Sentence weight also depends not only on the number of keywords present in it but also the weight of each keyword.
- **Upper Case Feature** - Sentences containing upper case words have been given additional weight as it is probable that they may contain proper nouns.

Sentences with higher weight are taken as the relevant sentences for the summary and arranged in the order they appear in the document yielding the required summary. The rearrangement becomes a challenge in the case of multiple documents. In that case, sentences are kept at the position at which they appear in original document (initial/middle/final). This rearrangement technique provides fair results.

## 4. Implementation

The Multiple Document Summarization System is implemented in Java using JAMA (Java Matrix Package) and WVTool (Word Vector Tool) packages. JAMA is used to perform all the matrix operations as computing SVD, eigen vector, Backprojection, etc. WVTool is used to

generate the Statistical Vector Space Model taking input as the Multiple Document Set or a Single document based on user requirement.

## 5. Evaluation of Summarizer

The present section will focus on the accuracy of the proposed summarization method. The accuracy of the method was examined on both single as well as multiple document summaries:

### 5.1 Single Document summary

Text belonging to different areas was taken. Summaries to the same texts were made by sentence extractions by different people. Based on the set of the summaries, we ranked sentences of the texts.

We then carried out the summarization process using our algorithm, the Auto Summarizer in MS Word, and the Gnome Summarizer and compared their agreement on the extracted sentences with the human sentence extractions.

The results are given in the following table.

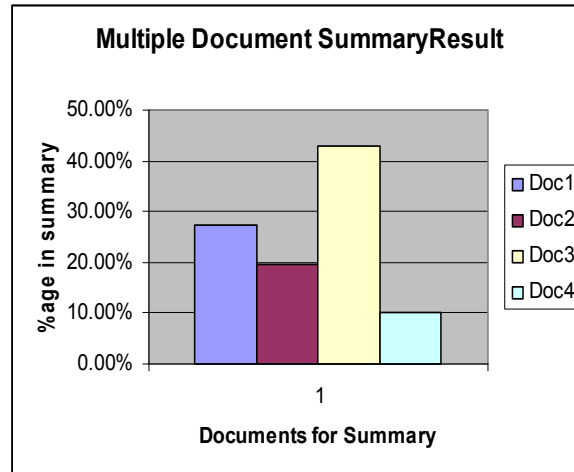
**Table 6. Summarization Algorithm Results**

Article #	Our Summarizer	MS Word Summarizer	Gnome Summarizer
Science (789 words)	60.0%	50.0%	70.0%
Geography (725 words)	55.56%	33.33%	22.5%
History (557 words)	70.0%	50.0%	48.5%
<b>Average accuracy</b>	<b>61.85%</b>	<b>44.44%</b>	<b>47%</b>

On an average, we get an average accuracy of 61.85% and improvement of 39.17% with respect to MS Word Summarizer.

### 5.2 Multiple Documents Summary

The set of Documents belonging to “Introduction to Web crawler” were taken and then summary was generated using the proposed algorithm, and it was observed that the summary thus generated was in coherence with most of the documents. The input documents set consisting of documents related to the topic for summarization has been shown in Table 7.



*Figure 2. Contribution percentage*

*Table 7. Input Set for Multiple Documents Summary*

Doc No.	Title of Doc	Doc Length
Doc1	Introduction to Crawler Architecture	1076 words
Doc2	Developing Web Search Engine	890 words
Doc3	Overview of Web Crawler	945 words
Doc4	Future of Search Engines	970 words

The cause of the low contribution of Doc4 to the summary generated was observed to be sentences with fewer keywords in them with respect to sentences from other documents, resulting in a low score of sentence.

## 6. Conclusion and Future Work

As seen from the results, the proposed method works better for various domains and, by using Semantic VSM instead of Statistical VSM; the summary obtained has become more informational and meaningful. Moreover, this method can be used to generate single as well as multiple document summaries.

The following areas in Multiple Document Summarization System require improvement:

1. Rearrangement of Extracted Sentences in the case of Multiple Documents Summarization to form an effective summary.
2. Enhance Flexibility of the system to generate a summary of multiple documents not necessarily belonging to the same topic.
3. Develop better methodology to incorporate the ACTION word score into Statistical VSM.
4. Evaluation of the system on large data samples.

## References

- Barzilay, R., "Information fusion in the context of multi-document summarization," Phd. Thesis, Columbia University, 2003.
- Bellare, K., A. Das Sarma, A. Das Sarma, N. Loival, V. Mehta, G. Ramakrishnan, and P. Bhattacharya, Generic Text Summarization using WordNet, <http://i.stanford.edu/~anishds/publications/lrec04/lrec04.ps>.
- Berry, M. W., S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information-retrieval," *Siam Review*, 37, 1995, pp. 573-95.
- Golub, G., and C. Van Loan, *Matrix Computations*, Baltimore: Johns Hopkins Univ Press, 1996.
- Gong, Y., and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," *SIGIR 2001*, pp. 19-25.
- Jessup, E. R., and D. C. Sorensen, "A parallel algorithm for computing the singular-value decomposition of a matrix," *Siam Journal on Matrix Analysis and Applications*, 15, 1994, pp. 530-548.
- Jolliffe, I. T., *Principal Component Analysis*, New York: Springer, 1986.
- Kupiec, J., J. Pedersen, and F. Chen, "A Trainable Document Summarizer," In *Proceedings of the 18th ACM-SIGIR Conference*, 1995, pp. 68-73.
- Moon, H., and P. J. Phillips, "Computational and Performance aspects of PCA-based Face Recognition Algorithms," *Perception*, 30, 2001, pp. 303-321.
- Otterbacher, J. C., A. J. Winkel, and D. R. Radev, The Michigan Single and Multidocument Summarizer for DUC 2002, [http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/umich\\_otter.pdf](http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/umich_otter.pdf).
- Pentland, A., B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21-23 June, 1994, Seattle, Washington, USA, pp. 84-91.
- Radev, D. R., and K. R. McKeown, "Generating natural language summaries from multiple on-line sources," *Computational Linguistics*, 24(3), 1998, pp. 469-500.

- Radev, D. R., H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility based evaluation, and user studies," In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April 2000.
- Radev, D., S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Celebi, H. Qi, D. Liu, and E. Drabek, "Evaluation challenges in large-scale multidocument summarization: the MEAD project," Johns Hopkins University CLSP Workshop Final Report, 2001.
- Radev, D. R., H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, 40, 2004, pp. 919-38.
- Salton, G., *The Smart Retrieval System*, Prentice Hall, Englewood Cliffs, N.J. 1971.
- Strang, G., *Introduction to Linear Algebra*, Wellesley, MA: Wellesley Cambridge Press, 1998.
- Turk, M., and A. Pentland, "Eigenfaces for Face Detection/Recognition," *Journal of Cognitive Neuroscience*, 3(1), 1991, pp. 71-86.
- Wall, M. E., A. Rechtsteiner, and L. M. Rocha, "Singular Value Decomposition and Principal Component Analysis," In *A Practical Approach to Microarray Data Analysis* (D.P. Berrar, W. Dubitzky, M. Granzow, eds.) Kluwer: Norwell, MA, 2003, pp. 91-109, LANL LA-UR-02-4001.

**Appendix A: Seed Word List**

Abstracted	Budgeted	Counseled	Enforced
Achieved	Built	Created	Enlightened
Acquired	Calculated	Critiqued	Enlisted
Acted	Cared	Cultivated	Ensured
Adapted	Charged	Dealt	Established
Addressed	Chartered	Debated	Estimated
Administered	Checked	Decided	Evaluated
Advertised	Clarified	Defined	Examined
Advised	Classified	Delegated	Exceeded
Advocated	Coached	Delivered	Excelled
Aided	Collaborated	Destruction	Expanded
Allocated	Collected	Designed	Expedited
Analyzed	Comforted	Detected	Experimented
Answered	Communicate	Determined	Explained
Anticipated	Compared	Developed	Explored
Applied	Completed	Devised	Expressed
Appraised	Complied	Diagnosed	Extracted
Approved	Composed	Directed	Facilitate
Arranged	Computed	Discovered	Fashioned
Ascertained	Conceived	Discriminated	Financed
Assembled	Conducted	Dispatched	Fixed
Assessed	Conserved	Displayed	Followed
Assisted	Consulted	Dissected	Formulated
Attained	Contracted	Documented	Fostered
Audited	Contributed	Drafted	Founded
Augmented	Converted	Drove	Gained
Authored	Cooperated	Edited	Gathered
Bolstered	Coordinated	Eliminated	Gave
Briefed	Copied	Empathized	Generated
Brought	Correlated	Enabled	Governed

*Principal Component Analysis Incorporating Semantic Vector Space Model*

Guided	Lifted	Perceived	Reduced
Handled	Listened	Perfected	Referred
Headed	Located	Performed	Related
Helped	Logged	Persuaded	Relied
Identified	Made	Planned	Reported
Illustrated	Maintained	Practiced	Researched
Imagined	Managed	Predicted	Responded
Implemented	Manipulated	Prepared	Restored
Improved	Mapped	Presented	Revamped
Improvised	Mastered	Prioritized	Reviewed
Inaugurated	Maximized	Produced	Scanned
Increased	Mediated	Programmed	Scheduled
Indexed	Memorized	Projected	Schemed
Indicated	Mentored	Promoted	Screened
Influenced	Met	Proposed	Set goals
Initiated	Minimized	Protected	Shaped
Inspected	Modeled	Proved	Skilled
Instituted	Modified	Provided	Solicited
Integrated	Monitored	Publicized	Solved
Interpreted	Narrated	Published	Specialized
Interviewed	Negotiated	Purchased	Spoke
Introduced	Observed	Queried	Stimulated
Invented	Obtained	Questioned	Strategized
Inventoried	Offered	Raised	Streamlined
Investigated	Operated	Ran	Strengthened
Judged	Ordered	Ranked	Stressed
Kept	Organized	Rationalized	Studied
Launched	Originated	Read	Substantiated
Learned	Overcame	Reasoned	Succeeded
Lectured	Oversaw	Recorded	Summarized
Led	Participated	Received	Synthesized

Supervised  
Supported  
Surveyed  
Sustained  
Symbolized  
Tabulated  
Talked  
Taught  
Theorized  
Trained  
Translated  
Upgraded  
Utilized  
Validated  
Verified  
Visualized  
Won  
Wrote



# A Study on Consistency Checking Method of Part-Of-Speech Tagging for Chinese Corpora<sup>1</sup>

Hu Zhang\* and Jiaheng Zheng\*

## Abstract

Ensuring consistency of Part-Of-Speech (POS) tagging plays an important role in the construction of high-quality Chinese corpora. After having analyzed the POS tagging of multi-category words in large-scale corpora, we propose a novel classification-based consistency checking method of POS tagging in this paper. Our method builds a vector model of the context of multi-category words along with using the  $k$ -NN algorithm to classify context vectors constructed from POS tagging sequences and to judge their consistency. These methods are evaluated on our 1.5M-word corpus. The experimental results indicate that the proposed method is feasible and effective.

**Keywords:** Multi-Category Words, Consistency Checking, Part of Speech Tagging, Chinese Corpus, Classification

## 1. Introduction

The construction of high-quality and large-scale corpora has always been a fundamental research area in the field of Chinese natural language processing. In recent years, rapid developments in the fields of machine translation (MT), information retrieval (IR), etc. are demanding more Chinese corpora of higher quality and larger scale. Ensuring the consistency of Part-of-Speech (POS) tagging plays an important role in the construction of high-quality Chinese corpora. In particular, we focus on consistency checking of the POS tagging of multi-tagged words, which consist of same Chinese characters and are nearly synonymous, yet have different grammatical functions. No matter how many different POS tags a multi-category word may be tagged with, it must be assigned the same POS tag when it appears in a similar context.

---

<sup>1</sup> This research was partially supported by the National Natural Science Foundation of China No. 60473139, 60775041 and the Natural Science Foundation of Shanxi Province No. 20051034.

\* School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China  
Tel: +86-3517010566  
E-mail: {zhanghu; jhzheng}@sxu.edu.cn

The general process of tagging text in the Chinese corpora consists of two steps: (1) automatic POS tagging is used to generate preliminary POS tags; (2) the preliminary POS tags are checked manually by domain experts. Novel approaches and techniques have been proposed for automatic rule-based and statistics-based POS tagging, and the “state-of-the-art” approaches achieve a tagging precision of at least 95% [Zhang *et al.* 1998; Huang *et al.* 2004; Xue *et al.* 2002; Ng *et al.* 2004]. A vast proportion of the words appearing in Chinese corpora are multi-category words. We have studied the textual data from the 2M-word Chinese corpus published by Peking University, and statistics show that the number of multi-category words is 5,473, covering 11% of the word types, while the number of word tokens is as high as 438,624, the percentage of which is 47%. Relevant works indicate that the different POS tag sets have different Stat. Huang *et al.* [2004] provided some important facts based on the 5M-word Sinica Corpus, wherein categorically ambiguous words only take up only 4.298% of all lexical items; however, categorically ambiguous words compose over 54.5% of all tokens. When checking the POS tags, human experts may have disagreements or may make mistakes in some cases. After analyzing 10,000 sentences containing the multi-category words, which were extracted from the 2M-word Chinese corpus of Peking University, the number of incorrect tags for multi-category words was found to be 133, which accounts for around 1.33% of the total.

So far in the field of POS tagging, most of the works have focused on novel algorithms or techniques for POS tagging. There are only a limited number of studies that have focused on consistency checking of POS tagging. Xing *et al.* [1999] analyzed the inconsistency phenomena of word segmentation (WS) and POS tagging. Qu and Chen [2003] improved the corpus quality by obtaining POS tagging knowledge from processed corpora. Qian and Zheng [2003] introduced a rule-based consistency checking method that obtained POS tagging knowledge automatically from processed corpora by machine learning (ML) and rough set (RS) methods. For real corpora, Du and Zheng [2001] proposed a rule-based consistency checking method to identify the inconsistency phenomena of POS tagging. However, the algorithms and techniques for automatic consistency checking of POS tagging proposed in the prior researches still have some insufficiencies. For example, the assignment of POS tags of the inconsistent POS tagging that are not included in the instance set needs to be conducted manually.

In this paper, we propose a novel classification-based method to check the consistency of POS tagging. Compared to Zhang *et al.* [2004], the proposed method fully considers the mutual relationships of the POS in POS tagging sequence, and adopts transition probability and emission probability to describe the mutual dependencies and the  $k$ -NN algorithm to weight the similarity. We evaluated our proposed algorithm on our 1.5M-word corpus. In an open test, our method achieved a precision rate of 85.24% and a recall rate of 85.84%.

The rest of the paper is organized as follows. Section 2 introduces the context vector model of POS tagging sequences. Section 3 describes the proposed classification-based consistency checking algorithm. Section 4 discusses the experimental results. Finally, the concluding remarks are given in Section 5.

## **2. Describing the Context of Multi-Category Words**

The basic idea of our approach is to use the contextual information of multi-category words to judge whether they are tagged consistently or not. In other words, if a multi-category word appears in two locations and the surrounding words in those two locations are tagged similarly, the multi-category word should be assigned with the same POS tag in those two locations as well. Hence, our approach is based on the context of multi-category words, and we model the context by looking at a window around a given multi-category word and the tagging sequence of this window. In the rest of this section, we describe the vector representation of the context of multi-category words and how to determine various parameters in our vector representations.

### **2.1 Vector Representation of the Context of Multi-Category Words**

Our vector representation of context consists of three key components: the POS tags of each word in a context window (*POS attribute*), the importance of each word to the center multi-category word based on distance (*position attribute*), and the dependency of POS tags of the center multi-category word and its surrounding words (*Dependency Attribute*).

Given a multi-category word and its context window of size  $l$ , we represent the words in sequential order as  $(w_1, w_2, \dots, w_l)$  and the POS tags of each word as  $(t_1, t_2, \dots, t_l)$ . We also refer to the latter vector as *POS tagging sequence*.

In practice, we choose a proper value of  $l$  so that the context window contains a sufficient number of words and the complexity of our algorithm remains relatively low. We will discuss this matter in detail later. In this study, we set the value of  $l$  to be 7, which means  $w_4$  is our multi-category word of interest and the context window includes 3 preceding and following words, where  $w_{4-i}$  is the  $i$ th preceding word and  $w_{4+i}$  is the  $i$ th following word.

#### **POS Attribute**

The POS tagging sequence contains information of the POS of each preceding and following word in a POS tagging sequence as well as the position of each POS tag. The POS of surrounding words may have different effect on determining the POS of the multi-category word, which we refer to as *POS attribute* and represent it using a matrix as follows.

**Definition 1.** *Suppose we have a tag set of size  $m(c_1, c_2, \dots, c_m)$ . Given a multi-category word with a context window of size  $l(w_1, w_2, \dots, w_l)$  and its POS tagging sequence  $(t_1, t_2, \dots, t_l)$ , the POS*

attribute matrix  $Y$  is an  $l$  by  $m$  matrix, where the rows indicate the POS tags of the preceding words, the multi-category word, and the following words in the context window, while the columns present tags in the tag set.  $Y_{ij}=1$  iff the POS tag of  $t_i$  is  $c_j$ .

For example, consider the POS attribute matrix of “比较” in the following sentence:

我们/r 就/d 能/v 比较/d 准确/a 地/u 预测/v 出/v 队员/n 的/u 成绩/n

As we let  $l=7$ , we look at the word “比较” and its 3 preceding and following words. Hence, the POS tagging sequence is  $(r,d,v,d,a,u,v)$ . In our study, we used a standard tag set that consists of 25 tags. Suppose the tag set is  $(n, v, a, d, u, p, r, m, q, c, w, l, f, s, t, b, z, e, o, l, j, h, k, g, y)$ , then the POS attribute matrix of “比较” in this example is:

$$Y = \begin{pmatrix} 0,0,0,0,0,0,1,\dots \\ 0,0,0,1,0,0,0,\dots \\ 0,1,0,0,0,0,0,\dots \\ 0,0,0,1,0,0,0,\dots \\ 0,0,1,0,0,0,0,\dots \\ 0,0,0,0,1,0,0,\dots \\ 0,1,0,0,0,0,0,\dots \end{pmatrix}$$

### Position Attribute

Due to the different distances from the multi-category word, the POS of the word before or after the multi-category word may, in a POS tagging sequence, have a different influence on the POS tagging of the multi-category word, which we refer to as *position attribute*.

**Definition2.** Given a multi-category word with a context window of size  $l$ , suppose the number of preceding or following words is  $n$  (i.e.,  $l=2n+1$ ), the position attribute vector  $V_x$  of the multi-category word is given by  $V_x=(d_1, \dots, d_n, d_{n+1}, \dots, d_l)$  where  $d_{n+1}$  is the value of the position attribute of the multi-category word and  $d_{n+1-i}(d_{n+1+i})$  is the value of the position attribute of the  $i$ th preceding (following) word. We further require that  $d_{n+1-i} = d_{n+1+i}$  and  $d_{n+1} + \sum_{i=1}^n (d_{n+1-i} + d_{n+1+i}) = 1$ .

We choose a proper position attribute vector so that the multi-category word itself has the highest weight, and the closer the surrounding word, the higher its weight is. If we consider a context window of size 7, based on our preliminary experiments, we chose the following position attribute values:  $d_1 = d_7 = 1/22$ ;  $d_2 = d_6 = 1/11$ ;  $d_3 = d_5 = 2/11$ ; and  $d_4 = 4/11$ . Hence, the final position attribute vector used in our study can be written as follows:

$$V_x = ((1/22), (1/11), (2/11), (4/11), (2/11), (1/11), (1/22)).$$

In the same way, if the size of a context window is 15, the position attribute vector can be described as follows:

$$V_X = ((1/190), (2/190), (4/190), (8/190), (16/190), (32/190), \\ ((64/190), (32/190), (16/190), (8/190), (4/190), (2/190), (1/190))$$

Note that, if the POS tag in the POS tagging sequence is incorrect, the position attribute value of the corresponding position should be turned into a negative value, so that when the incorrect POS tag appears in a POS tagging sequence, this attribute can correctly show that the incorrect POS tag has had a negative effect on generating the final context vector.

### **Dependency Attribute**

The last attribute we focus on is *dependency attribute*, which corresponds to the fact that there are mutual dependencies on the appearance of every POS in POS tagging sequences. In particular, we use *transition probability* and *emission probability* in Hidden Markov Model (HMM) to capture this dependency.

**Definition 3.** Given a tag set of size  $m(c_1, c_2, \dots, c_m)$ , the transition probability table  $T$  is an  $m$  by  $m$  matrix and given by:

$$T_{i,j} = P^T(c_i, c_j) = \frac{f(c_i, c_j)}{f(c_i)}$$

where  $f(c_i, c_j)$  is the frequency of the POS tag  $c_j$  appears after the POS tag  $c_i$  in the entire corpus;  $f(c_i)$  is the frequency of the POS tag  $c_i$  appearing in the entire corpus; and  $P^T$  is the transition probability.

**Definition 4.** Given a tag set of size  $m(c_1, c_2, \dots, c_m)$ , the emission probability table  $E$  is an  $m$  by  $m$  matrix and given by:

$$E_{i,j} = P^E(c_i, c_j) = \frac{f(c_i, c_j)}{f(c_j)}$$

where  $f(c_i, c_j)$  is the frequency of the POS tag  $c_i$  appearing before the POS tag  $c_j$  in the entire corpus;  $f(c_j)$  is the frequency of POS tag  $c_j$  appearing in the entire corpus; and  $P^E$  is the emission probability.

Note that both  $T$  and  $E$  are constructed from the entire corpus and that we can look up these two tables easily when we consider the POS tags appears in POS tagging sequences.

Now, when we look at a context window of size 7 ( $w_1, w_2, \dots, w_7$ ) and its POS tagging sequence ( $t_1, t_2, \dots, t_7$ ), there are three types of probabilities we need to take into account.

The first one is the probability of the appearance of the POS tag  $t_4$  of the multi-category word, which we can write as follows:

$$P^{CX}(t_4) = f(\omega_4 \text{ is tagged as } t_4) / f(\omega_4)$$

where  $f(w_d)$  is the frequency of the appearance of the multi-category word  $w_d$  in the entire corpus and  $f(w_d \text{ is tagged as } t_d)$  is the frequency of the appearance where the word  $w_d$  is tagged as  $t_d$  in the entire corpus.

The second one is transition probability, which is the probability of the appearance of the POS tag  $t_{i+1}$  in the  $i+1$  position after the POS tag  $t_i$  in the  $i$  position and is shown as follows:

$$P_{(i,i+1)}^T = P^T(t_i, t_{i+1}) = f(t_i, t_{i+1})/f(t_i)$$

The last is emission probability, which is the probability of the appearance of the POS tag  $t_{i-1}$  in the  $i-1$  position before the POS tag  $t_i$  in the  $i$  position and is shown as follows:

$$P_{(i-1,i)}^E = P^E(t_{i-1}, t_i) = f(t_{i-1}, t_i)/f(t_i)$$

According to the above three probability formulas, we can build a seven-dimensional vector, where each dimension corresponds to one POS tag, respectively.

**Definition 5:** Given a multi-category word with a context window of size 7 and its POS tagging sequence, the dependency attribute vector  $V_P$  of the multi-category word is defined as follows:

$$V_P = (P_1, P_2, P_3, P_4, P_5, P_6, P_7),$$

where

$$P_1 = P_{(1,2)}^T \times P_2; P_2 = P_{(2,3)}^T \times P_3; P_3 = P_{(3,4)}^T \times P_4; P_4 = P^{CX}(t_4);$$

$$P_5 = P_{(4,5)}^E \times P_4; P_6 = P_{(5,6)}^E \times P_5; P_7 = P_{(6,7)}^E \times P_6;$$

### Context Vector of Multi-Category Words

Now we are ready to define the context vector of multi-category words.

**Definition 6.** Given a multi-category word with a context window of size  $l$  and its POS attribute matrix  $Y$ , position attribute vector  $V_X$ , and dependency attribute vector  $V_P$ , the context vector  $V_S$  of the multi-category word is defined as follows:

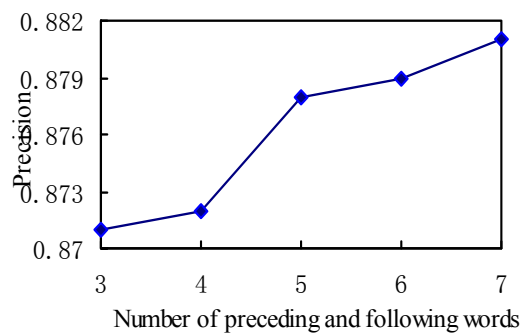
$$V_S = (\alpha V_X + \beta V_P) \times Y$$

where  $\alpha$  and  $\beta$  are the weights of the position attribute and the dependency attribute, respectively.

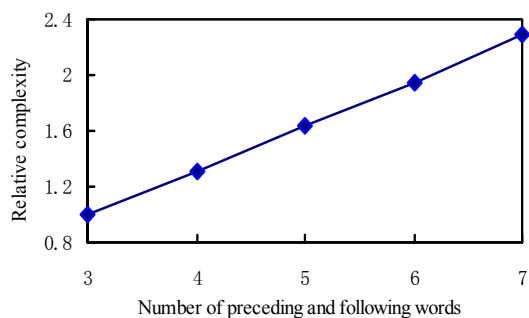
Note that we require  $\alpha + \beta = 1$ , and their optimal values are determined by experiments in our study.

## 2.2 Experiment on the Size of the Context Window

Context vectors can be extended by using 4 to 7 preceding and following words as a substitute for the 3 preceding and following words we used in context windows and POS tagging sequences. We conducted experiments with a context window of size 3 to 7 on our sampled 1M-word training corpus performing closed tests. The experimental results are simultaneously evaluated in terms of both the precision of consistency checking and algorithm complexity. We plot the effect on precision and relative complexity of the number of preceding or following words in Figure 1 and Figure 2, respectively.



**Figure 1.** *Effect on precision of the number of preceding and following words.*



**Figure 2.** *Effect on complexity of the number of preceding and following words.*

As shown in Figure 1, the precision of consistency checking increases as we include more preceding and following words. In particular, the precision is improved by 1% when we use 7 preceding and following words.

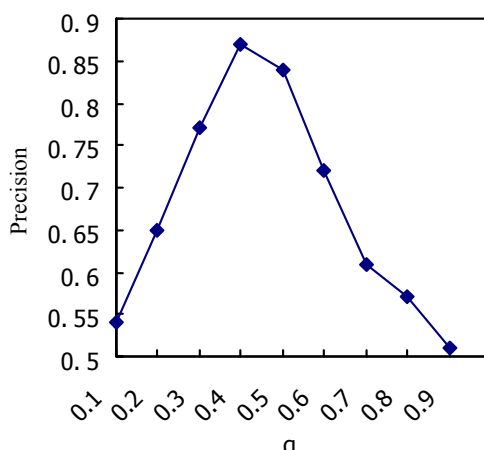
When the context vector model is extended to contain 7 preceding and following words, the dimensionality of the position attribute vector, POS attribute vector, and dependency

attribute vector increases from 7 to 15, which results in a significant change in run time. Suppose the run time of our algorithm with 3 preceding and following words is 1, then, extending the context vector to 15 dimensions, the change in run time of our algorithm is plotted in Figure 2.

As shown in Figures 1 and 2, the increase of complexity is much higher than that of precision when extending the number of preceding and following words. Hence, we chose 3 as the number of preceding and following words to form context windows and calculate context vectors.

### 2.3 Effect on Consistency Checking Precision of $\alpha$ and $\beta$

When using our sampled 1M-word training corpus to conduct closed test, we found that consistency checking precision changes significantly with different values of  $\alpha$  and  $\beta$ . Figure 3 shows the trend when  $\alpha$  varies from 0.1 to 0.9.



**Figure 3.** Effect on consistency checking precision of  $\alpha$  and  $\beta$ .

As shown in Figure 3, when  $\alpha=0.4$  ( $\beta=0.6$ ), consistency checking precision reaches the highest value. Hence, we used  $\alpha=0.4$  and  $\beta=0.6$  in our experiments. At the same time, the results show  $\alpha$  and  $\beta$  have an important effect on consistency checking precision, namely the position attribute vector and the dependency attribute vector are very important for describing the context vector of multi-category word, which is consistent with our experience.

### 3. Consistency Checking of POS Tagging

Our consistency checking algorithm is based on classification of context vectors of multi-category words. In particular, we first classify context vectors of each multi-category



word in the training corpus, and then we conduct the consistency checking of POS tagging based on classification results.

### 3.1 Similarity between Context Vectors of Multi-category Words

After constructing context vectors for all multi-category words from their context windows and POS tagging sequences, the similarity of two context vectors is defined as the *Euclidean Distance* between the two vectors.

$$d(x, y) = \|x - y\| = \left[ \sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$$

where  $x$  and  $y$  are two arbitrary context vectors of  $n$  dimensions.

### 3.2 $k$ -NN Classification Algorithm

Classification is a process to assign objects that need to be sorted into certain classes. In this paper, we used a popular classification method: the  $k$ -NN algorithm.

Suppose we have  $c$  classes and a class ( $\theta_i (i=1, 2, \dots, c)$ ) has  $N_i$  samples ( $x_j^{(i)} (j=1, 2, \dots, N_i)$ ). The idea of the  $k$ -NN algorithm is that for each unlabeled object  $x$ , compute the distances between  $x$  and  $M = \sum_{i=1}^c N_i$  samples ( $x_j^{(i)}$ ) whose class is known, and select  $k$  samples ( $k$  nearest neighbors) with the smallest distance. This object  $x$  will be assigned to the class that contains the most samples in the  $k$  nearest neighbors.

We now formally define the discriminant function and discriminant rule. Suppose  $k_1, k_2, \dots, k_c$  are the numbers of samples in the  $k$  nearest neighbors of the object  $x$  that belong to the classes  $\theta_1, \theta_2, \dots, \theta_c$ , respectively. Define the discriminant function of the class  $\theta_i$  as  $d_i(x) = k_i, i = 1, 2, \dots, c$ . Then, the discriminant rule determining the class of the object  $x$  can be defined as follows:

$$d_m(x) = \max_{i=1, 2, \dots, c} d_i(x) \Rightarrow x \in \theta_m$$

### 3.3 Consistency Checking Algorithm

In this section, we describe the steps of our classification-based consistency checking algorithm in detail.

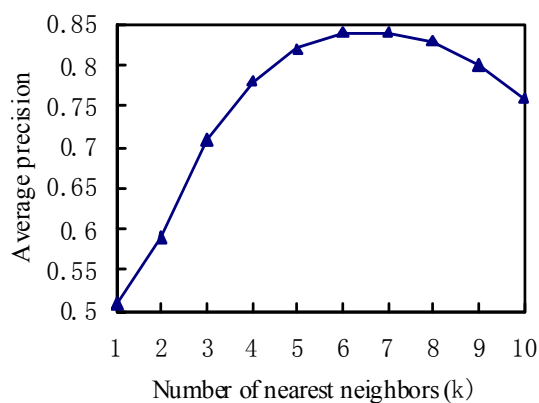
- Step 1: Randomly sampling sentences containing multi-category words and checking their POS tagging manually. For each multi-category word, classifying the context vectors of the sampled POS tagging sequences, so that the context vectors that have the same POS for the multi-category word belonging to the same class. Let  $C_{ci}$  ("ci" is a multi-category word) be the number of possible POS tags of the multi-category word  $ci$ .

- Step 2: Give a context vector  $v$  of a multi-category word  $ci$ , calculating the distances between  $v$  and all the context vectors that contain the multi-category word  $ci$  in the training corpus, and selecting  $k$  context vectors whose distances are smaller than the others.
- Step 3: According to the  $k$ -NN algorithm, checking the classes of the  $k$  nearest context vectors and classifying the vector  $v$ .
- Step 4: Comparing the POS of the multi-category word  $ci$  in the class that the  $k$ -NN algorithm assigns  $v$  to and the POS tag of  $ci$ . If they are the same, the POS tagging of the multi-category word  $ci$  is considered to be consistent, otherwise it is inconsistent.

The major disadvantage of this algorithm is the difficulty in selecting the value of  $k$ . If  $k$  is too small, the classification result is unstable; on the other hand, if  $k$  is too big, the classification deviation increases. Therefore, the selection of the value of  $k$  should depend on not only some general factors, but also the unique characteristics and structures of the data set. Hence, the best method of selecting  $k$  should be a strategy that has an ability to adapt different data sets. One such strategy is to try different values on the training set first, draw the performance curve, and select an optimal  $k$  value.

### 3.4 Selecting $k$ in Classification Algorithm

Figure 4 shows the consistency checking precision values obtained with various  $k$  values in the  $k$ -NN algorithm. The precision values are closed test results on our 1M-word training corpus, and were obtained by using  $\alpha=0.4$  and  $\beta=0.6$  in the context vector model.



**Figure 4. Effect on precision of  $k$  in the  $k$ -NN algorithm.**

As shown in Figure 4, when  $k$  continues to increase from 6, the precision remains the same. When  $k$  reaches 9, the precision starts declining. Our experiment with other  $\alpha$  and  $\beta$  values also show similar trends. Hence, we chose  $k=6$  in this paper.

#### 4. Experimental Results and Discussions

We evaluated our consistency checking algorithm on our 1.5M-word corpus (including the 1M-word training corpus) and conducted open and closed tests. The results are showed in Table 1.

**Table 1. Experimental Results.**

Test corpora	Test type	Number of multi-category words	Number of the true inconsistencies	Number of the identified inconsistencies (True)	Recall (%)	Precision (%)
1M word	closed	127,210	1,147	1,219(1063)	92.67	87.20
500K word	open	64,467	579	583(497)	85.84	85.24

$$\text{Recall} = \frac{\text{Number of the identified true inconsistencies}}{\text{Number of the true inconsistencies}}$$

$$\text{Precision} = \frac{\text{Number of the identified true inconsistencies}}{\text{Number of the identified inconsistencies}}$$

The experimental results show two interesting trends. First, the precision and recall of our consistency checking algorithm are 87.20% and 92.67% in the closed test, respectively, and 85.24% and 85.84% in the open test, respectively. Compared to Zhang *et al.* 2004, the precision of consistency checking is improved by 2~3%, and the recall is improved by 10%. The experimental results indicate that the context vector model has great improvements over the one used in Zhang *et al.* 2004. Second, thanks to the considerable improvement of the recall, to some extent, our consistency checking algorithm prevents the occurrence of events with small probabilities in POS tagging.

#### 5. Conclusion and Future Research

After analyzing the experimental results we have come to the following conclusions:

- The proposed classification-based method is effective for consistency checking of POS tagging. It solves a difficult problem that easily appears in automatic POS tagging processes, and greatly improves the efficiency and quality of manual correction of automatic POS tagging.
- The information in corpora can be considered as if it were in a multi-dimensional space with various attributes that have mutual influences. When conducting a consistency checking of POS tagging, it is incomplete if we only consider POS tagging and relationships between the POS without taking into account other attributes of words. We

may use other widely known attributes of words to improve the consistency checking of POS tagging.

In the future, we plan to investigate more types of word attributes and incorporate linguistic and mathematical knowledge to develop better consistency checking models, which ultimately provide a better means of building high-quality Chinese corpora.

### Acknowledgements

This research was partially supported by the National Natural Science Foundation of China No. 60473139, 60775041 and the Natural Science Foundation of Shanxi Province No. 20051034.

### References

- Du, Y.-P., and J.-H. Zheng, "The proofreading method study on consistence of segment and part-of-speech," *Computer Development and Application*, 14(10), 2001, pp. 16-18.
- Huang, C.-R., and R.-Y. Chang, "Categorical Am biguity and Information Content: A Corpus-based Study of Chinese," *Journal of Chinese Language and Computing*, 14(2), 2004, pp. 157-165.
- Ng, H. T., and K. L. Jin, "Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?," *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, 2004, Barcelona, Spain, pp. 277-284.
- Qian, W.-N., and A.-Y. Zhou, "Analyzing Popular Clustering Algorithms from Different Viewpoints," *Journal of Software*, 13(8), 2002, pp. 1382-1394.
- Qian, Y.-L., and J.-H. Zheng, "Research on the method of automatic correction of chinese POS tagging," *Journal Of Chinese Information Processing*, 18(2), 2003, pp. 30-35.
- Qian, Y.-L., and J.-H. Zheng, "An approach to improving the quality of part-of-speech tagging of Chinese text," *In Proceedings of the 2004 IEEE International Conference on Information Technology: Coding and Computing (ITCC 2004)*, 2004, USA, pp. 183-187.
- Qu, W.-G., and X.-H. Chen, "Research and practice on the machine proofreading of the tagged corpora," *In Proceeding of the 7<sup>th</sup> National Computation Linguistics (JSCL'03)*, 2003, Beijing, China, pp. 318-324.
- Xing, H.-B., "Analysing on the words classified hard in pos tagging," *In Proceedings of the 5th National Computational Linguistics (JSCL'99)*, 1999, Beijing, China, pp. 187-192.
- Xue, N., F.-D. Chiou, and M. Palmer, "Building a Large-Scale Annotated Chinese Corpus," *In Proceedings of the 19th international conference on Computational linguistics*, 2002, Taipei, Taiwan, pp. 1-8.
- Zhang, H., and J.-H. Zheng, "The Inspecting Method Study on Consistency of POS Tagging of Corpus," *Journal Of Chinese Information Processing*, 18(5), 2004, pp. 11-16.

Zhang, M., and S. Li, "POS Tagging Chinese Corpus Based on Statistics and Rules," *Journal of Software*, 9(2), 1998, pp. 134-138.



# Constructing a Temporal Relation Tagged Corpus of Chinese Based on Dependency Structure Analysis

Yuchang CHENG\*, Masayuki ASAHARA\* and Yuji MATSUMOTO\*

## Abstract

This paper describes an annotation guideline for a temporal relation-tagged corpus of Chinese. Our goal is construction of corpora to be used for a corpus-based analysis of temporal relations among events. Since annotating all combinations of events is inefficient, we examine the use of dependency structure to efficiently recognize temporal relations. We annotate a part of Treebank based on our guidelines. Then, we survey a small tagged data set to investigate the coverage of our method. While we find that use of dependency structure drastically reduces manual effort in constructing a tagged corpus with temporal relations, the coverage of the methods achieves about 63%.

**Keywords:** Temporal Entities, Event Entities, Temporal Reasoning, Event Semantics, Dependency Structure

## 1. Introduction

Extracting temporal information in documents is a useful technique for many NLP applications such as question answering, text summarization, machine translation, and so on. Temporal information includes three elements: 1. temporal expressions, which describe time or period in the real or virtual world; 2. event or situation expressions that occur instantaneously or that last for a period of time; 3. temporal relations, which describe the ordering relation between an event expression and a temporal expression or between two event expressions. There is a great deal of research dealing with temporal expressions and event expressions. Extracting temporal expressions is a subtask of NER [IREX committee 1999] and is widely studied in many languages [Mani *et al.* 2006b]. Normalizing temporal expressions has been investigated in evaluation workshops [Chinchor 1997]. Event semantics has been investigated in linguistics and in AI fields [Bach 1986]. However, research on temporal relation extraction is still limited. Temporal relation extraction includes the following issues: identifying events, anchoring an event to the timeline, ordering events, and reasoning of

---

\* Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan  
E-mail: {yuchan-c, masayu-a, matsu}@is.naist.jp

contextually underspecified temporal expressions. To extract temporal relations, several knowledge sources are necessary, such as tense and aspect of verbs, temporal adverbs, and world knowledge [Mani *et al.* 2006b].

The goal of our research is to efficiently construct a temporal relation tagged corpus of Chinese. In English, TimeBank [Pustejovsky *et al.* 2006], a temporal information tagged corpus, is available for introducing machine learning approaches to automatically extract temporal relations. In Chinese, there is some related research on temporal expression extraction [Li *et al.* 2005]. However, there is no publicly available resource for temporal information processing in Chinese. Currently, such resources, event and temporal relation tagged corpora, are being made. Annotating all temporal relations of event pairs is time-consuming. Therefore, we propose a dependency structure based method to annotate temporal relations manually on a limited set of event pairs and extend the relations using inference rules. This method reduces manual effort. The dependency structure helps to detect subordinate and coordinate structures in sentences. We also describe a guideline for corpus annotation. Our annotation guideline is based on TimeML [Sauri *et al.* 2005], which was originally designed for English texts. We have developed a machine learning based dependency analyzer for Chinese [Cheng 2005]<sup>1</sup> and a Chinese morphological analyzer [GOH 2006]. We can use our dependency analyzer to analyze raw text then use the output to annotate the temporal relations. However, in this paper, we use a syntactic tagged Chinese treebank (Penn Chinese Treebank) [Palmer *et al.* 2005] to create a temporal information annotated corpus. Finally, we survey the distribution of the temporal relations in our tagged corpus. We also evaluate the coverage of limited event pairs in our criteria.

## 2. Background

We investigated the distribution of events and temporal expressions in TimeBank [Pustejovsky *et al.* 2006] and Penn Chinese Treebank. TimeBank is a temporal information tagged corpus of English with TimeML annotation guideline. We found that the distribution of events and temporal expressions is uneven. Therefore, our corpus does not focus on the relations between an event and a temporal expression, but between two events. TimeML is a corpus annotation guideline of temporal information for English news articles. In this section, first, we introduce the resource -- TimeBank and describe the temporal relation links of original TimeML. Second, we investigate the distribution of events and temporal expressions in TimeBank and Penn Chinese Treebank. Finally, we analyze the temporal relation links to observe the correlation between dependency structure and TimeML links.

---

<sup>1</sup> The dependency analyzer is trained on Penn Chinese Treebank. The accuracy of the head-modifier relation analysis is 89%.



**Table 1. Tags of TimeML annotation**

Tags	Definition
TimeML tags	
EVENT	Situations that “happen” or “occur”, includes tensed /untensed verbs, nominalizations, adjectives, predicative clauses or prepositional phrases
TIMEX3	Temporal expressions, includes date, time and duration.
SIGNAL	Textual elements that make explicit the relation holding between two entities
MAKEINSTANCE	To create the actual realizations of an event
Link tags	
TLINK	Temporal relation links, represents the temporal relationship holding between two temporal entities (TIMEX3 and event)
SLINK	Subordinate links, represents contexts introducing relations between two events
ALINK	Aspectual links, represents the relationship between an aspectual event and its argument event

## 2.1 Resources

TimeML is a corpus guideline of temporal information for English news articles. Table 1 lists the definition of the tags. “EVENT”, “TIMEX3” and “SIGNAL” tags in TimeML mark temporal entities such as event expressions and temporal expressions. Link tags annotate temporal relations between entities. The tag “TLINK” represents the temporal relationship between two entities. The definition of temporal relations using the tag “TLINK” is based on Allen’s temporal relations [Allen 1983]. The tags “SLINK” and “ALINK” annotate the relations between a main event and its subordinate event. While the tag “ALINK” describes an aspectual relation, the tag “SLINK” describes a subordinate relation without explicit aspectual meaning. TimeBank is a temporal information tagged corpus that includes full temporal information (temporal expressions, events and temporal relations). The corpus is annotated according to the TimeML guidelines.

We create a temporal information tagged corpus for Chinese with our criteria because there is no such tagged corpus currently published. Our criteria include many elements of dependency structure. Sentences should be parsed to dependency structures; then, one should use the information of dependency structure to annotate temporal relations. Therefore, we used Penn Chinese Treebank [Palmer *et al.* 2005] as the original data and transfer phrase structures to dependency structures (See Section 4.1).

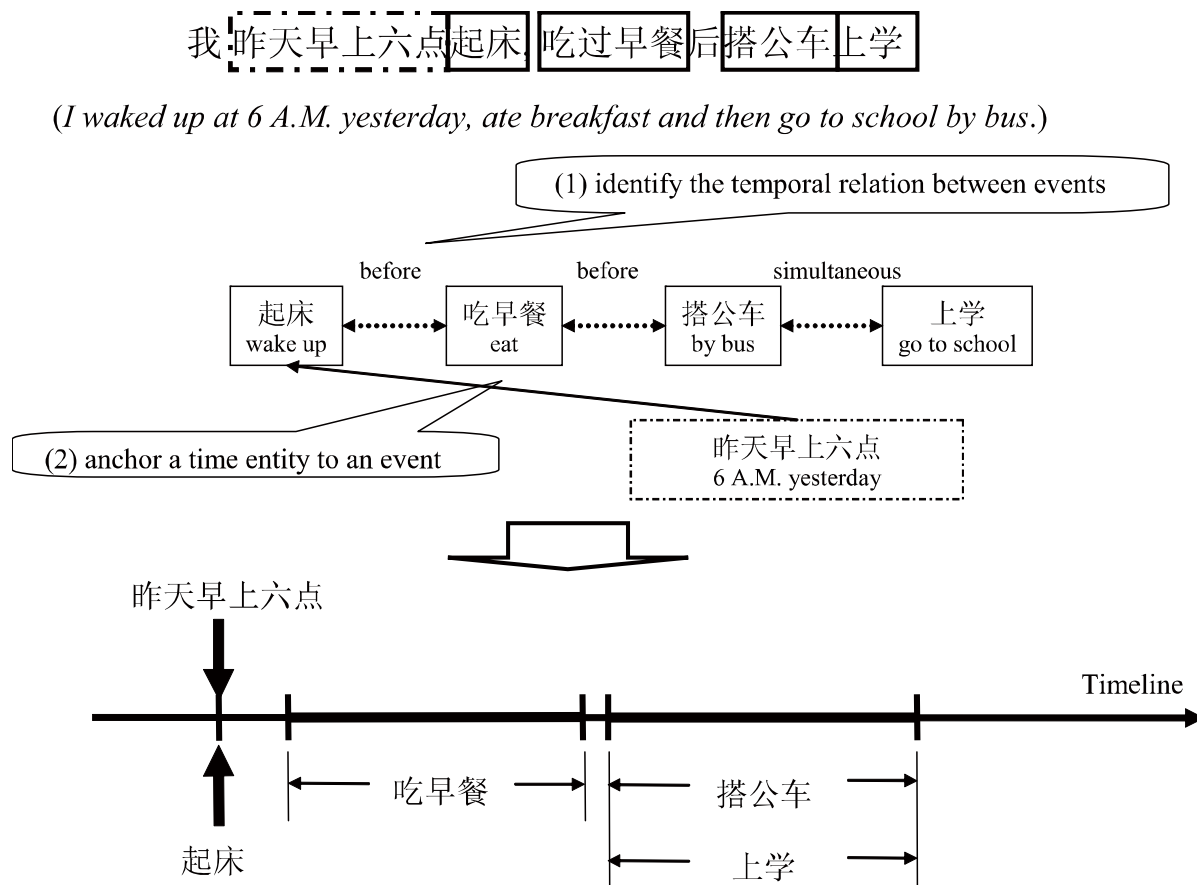


Figure 1. The temporal relations between events and a temporal expression

## 2.2 The Temporal Relations: “between a temporal expression and an event” OR “between two events”

Temporal relation includes anchoring a relation from an event to a temporal expression and ordering the relations between two events. Intuitively, ordering two events requires a temporal expression that can anchor events on the timeline. Previous research [Li et al. 2001] focused on simple anchoring problems, such as co-reference resolution between an event and a temporal expression. However, recent research ([Mani et al. 2006b], [Li et al. 2004]) has covered temporal relations depending on not only the temporal expressions but also world knowledge such as verb ontology in VerbOcean [Chklovski and Pantel 2004] and inference rules.

However, some events cannot be anchored on a timeline without ordering the events independently. For example, in Figure 1, there is one temporal expression “昨天早上六点” (6 A.M. yesterday) and four verb phrases (“起床” (wake up), “吃早餐” (eat breakfast), “搭公车”

(by bus), and “上学” (go to school)) in the example sentence. For ordering these events (verbal phrases) on the timeline, we can analyze the temporal relationship between an event and a temporal expression. In this example, there is only one temporal expression “昨天晚上六点” (6 A.M. yesterday) that can be analyzed, and it is the anchor time of the event “起床” (wake up). Figure 1 (1) describes the temporal relations of adjacent event pairs<sup>2</sup>, and we can recognize some temporal relations by considering the event pairs rather than the temporal expression. These temporal relations are: the event “起床” (wake up) occurs before the event “吃早餐” (eat breakfast); the event “吃早餐” (eat breakfast) occurs before the event “搭公车” (by bus); and the events “搭公车” (by bus) and “上学” (go to school) occur at the same time. Figure 1 (2) describes the temporal relation between the anchor time “昨天晚上六点” (6 A.M. yesterday) and the event “起床” (wake up). This temporal relation can anchor the event on the timeline. Combining the temporal relations in Figure 1 (1) and (2), the reader can recognize what happened and when they happened. Therefore, we can divide the process of recognizing the temporal information in the sentence into two steps: (1) recognizing the temporal relation between two events; (2) anchoring the events on the timeline. In extreme cases, we can think that a reader can recognize the situation by only considering the temporal relation between events, even if the reader does not know the anchor time of the events. Therefore, we think that annotating the temporal relations between event pairs is an independent task in temporal information processing. In this research, we focus on annotating the relations between two events.

**Table 2. Distribution of tags in TimeBank**

Distribution of temporal entities tags					
Tags	EVENT	MAKEINSTANCE	TIMEX3	SIGNAL	
Number	7935	7940	1414	688	
Distribution of temporal links in adjacent and dependency structure viewpoints					
	Entities	all links	adjacent relations	head-modifier relations	adjacent and head-modifier relations
TLINK	Timex3 and event	6418	3467	1372	4458
	Event and event	3314	1757	1186	2826
SLINK	Event and event	2932	2129	2174	2833
ALINK	Event and event	265	167	157	251

Considering the distribution of the events in TimeBank (see Table 1 and Table 2), the number of events is more than the number of temporal expressions (TIMEX3). Therefore, to order the events, many events should share a temporal expression or should be analyzed as the temporal relation of event pairs with no corresponding temporal expression. We observe

<sup>2</sup> These temporal relations of the adjacent event pairs do not include all recognizable temporal relations.

similar things in Chinese corpora. We calculate the distribution of the temporal phrases and verbs in the Penn Chinese Treebank [Palmer *et al.* 2005]. There are 72,245 verbs (the words with POS-tags VV, VA, VE, VC) and 10,129 temporal phrases (the phrases with the tag “TMP”) in this Treebank. Only a portion of the verbs in an article have their own temporal expression (phrases), other verbs do not have direct temporal expression to anchor the verbs on the timeline. If we consider the verbs as the events in Treebank, most of the temporal relations are not between a temporal expression and an event, but between two events. To analyze the temporal relations between events that do not have their own temporal expression is necessary for recognizing the temporal information in Chinese articles.

### 2.3 Adjacent Links in TimeBank

TimeBank 1.2<sup>3</sup> contains 183 articles with over 61,000 non-punctuation tokens. We investigate the distribution of temporal tags as shown in Table 2. TimeBank includes 9,615 links (TLINK, SLINK, and ALINK), of which, 5,763 links are the relations between adjacent entity pairs<sup>4</sup> (an adjacent pair means the focal event and its linearly preceding event). According to the distribution, if we are able to recognize adjacent relations correctly (at least 60% (5,763/9,615)<sup>5</sup> of temporal relations are recognized), we expect to acquire more temporal relations with an additional process, such as adaptation of inference rules that we will describe in Section 4.4. We refer to the links of adjacent relations as “adjacent links”. To recognize the adjacent links of events, we annotate adjacent event pairs.

Additionally, we find that about 40% (2,296/5,763) of the links in the adjacent links are SLINKs and ALINKs. The links with the tag “SLINK” mean subordinate relations between events (not from an event to a temporal expression). Subordinate relation is the relationship between a focus event and a main event that the focus event depends on. SLINKs do not include TLINKs. If we extract SLINKs first, extracting other TLINKs from the remaining temporal entities would become simple. This observation gives us the idea that recognition of SLINKs is an important task for annotating adjacent relations.

---

<sup>3</sup> <http://www ldc.upenn.edu/>

<sup>4</sup> The tag “TLINK” includes the temporal relations between document creation time and other temporal entities in an article, and includes the temporal relations between two matrix verb events in different sentences.

<sup>5</sup> An adjacent link could simultaneously be a head-modifier link. Therefore the numbers of the column “adjacent links” and the column “head-modifier links” in Table 2 are not complementary. The column “all links” is not the sum of “adjacent links” and “head-modifier links”. Further more, we only consider the dependency structure of “sentences”. If the relation links in Timebank cross different sentences, the dependency structures cannot recognize these links. The remnants of SLINKs that are not head-modifier relations are the links crossing different sentences.

## **2.4 Links on the Dependency Structure in TimeBank**

Since the majority of adjacent links are subordinate relations, we cannot analyze the temporal relations between contiguous pairs of matrix verb events without analyzing the structure of the subordinate relations, namely dependency structure. To calculate the distribution of links in TimeBank using dependency structure, we parse the sentences in TimeBank into the dependency structure and estimate the number of head-modifier (governor-dependent) relations that are SLINK or ALINK. We use the POS-tagger “TnT” [Brants 2000] to tag the sentences and use the MST parser<sup>6</sup> [McDonald *et al.* 2005] to parse sentences to dependency structures. The column of “Head-modifier links” in Table 2 shows the number of each type of link that is a head-modifier relation. Seventy-three (2,331/3197) percent of S/ALINKs (SLINK + ALINK) in TimeBank are of head-modifier relations. This shows that dependency structure can be used to extract most S/ALINKs in English articles. Note that an event pair in a head-modifier relation link can also be one in an adjacent relation link.

Previous researchers have shown that syntactic information is useful for temporal information extraction [Li *et al.* 2004]. We use dependency structure for annotating temporal relations. The reason is that dependency structures are simpler and more comprehensible than phrase structures. The dependency grammar is composed of asymmetric head-modifier relations between words. We focus on the relation of event pairs. Dependency structure can describe the semantic relations between events clearly. Subordinate relations can be identified by the dependency structure. Therefore, dependency structure analysis is very useful for annotating the temporal relation.

## **3. Strategy of Chinese Temporal Information Annotation**

We propose an annotation guideline for developing a Chinese temporal relation tagged corpus. The guideline is based on TimeML. TimeBank includes all possible temporal relations between two entities and is annotated manually. However, to annotate full temporal information of a newswire text requires considerable human effort and cost. To reduce human effort, we introduce several constraints on the original TimeML. First, we limit the definition of events to verbs. Second, we focus on three types of event pairs in a complete graph according to dependency structure and use inference rules to extend relations.

### **3.1 The Definition of the Events**

First, we limit the definition of events to verbs. According to the TimeML guideline for English, verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases can be events. However, recognizing an instance whether nominalization represents an event or

---

<sup>6</sup> We train the MST parser using Penn Treebank.

not is difficult in Chinese articles. For example, a nominalization “电话(*telephone*)” could mean a telephone machine in the example “我买了一具电话(*I bought a telephone*)”, or could mean a telephone call in the example “他在电话中说...(In the telephone call, he said...)”. Similar to the example, the semantic role of most of the aforementioned non-finite entities (nominalizations and adjectives) is ambiguous in the morphological analysis and the meaning of an entity depends on the context. In other types of events (predicative clauses and prepositional phrases), recognizing these entities from a context needs chunking techniques. The phrases / clauses usually have the hierarchical structure of verbs. It will be complicated to recognize these event entities when we extract the events automatically. It is also difficult to recognize events from all of the aforementioned entities except for verbs. Therefore, to simplify the process of recognizing events, we only regard verbs as events in our research.

It should be noted that we do not limit the domain of verbs. In the related research [Li et al. 2001], the researchers manually created a dictionary which included the common verbs in Chinese financial news articles and recognized the event using the dictionary. However, our original data do not limit the domain of articles. We should consider all of the verbs in corpus to assure the multiplicity of our corpus.

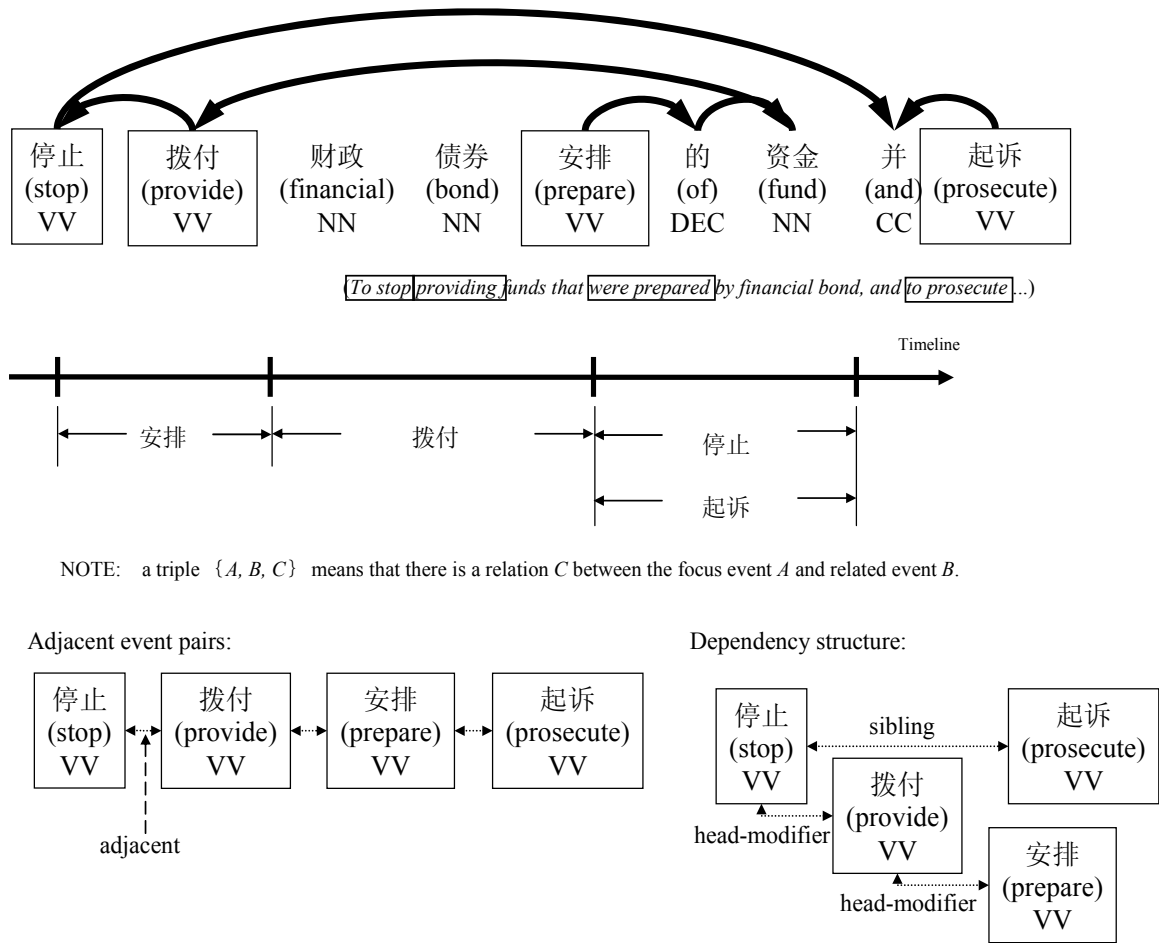
### 3.2 Three Types of Event Pairs

Second, we focus on three types of event pairs in the complete graph. The first type is adjacent event pairs. The second and third types are the head-modifier event pairs and the sibling event pairs in dependency tree representation of a sentence. The first type (adjacent event pairs as seen in Section 2.3) and the other two types (head-modifier or sibling event pairs as seen in Section 2.4) are not exclusive. According to our investigation in TimeBank, subordinate event pairs are head-modifier relations and coordinate event pairs are sibling relations. Therefore, using dependency structure can extract subordinate relations and coordinate relations in a sentence.

The three types of pairs are shown in Figure 2. The example phrase “停止拨付财政债券安排的资金并起诉 (*To stop providing funds that were prepared by financial bond, and to prosecute...*)” in Figure 2 has four events: “停止 (*stop*)”, “拨付 (*provide*)”, “安排 (*prepare*)” and “起诉 (*prosecute*)”. The temporal relations of all possible event pairs are shown in the row “All possible temporal relations” of the table in Figure 2. For example, the temporal relation: {安排, 拨付, before}, means that the event “安排(*prepare*)” occurs before the event “拨付(*provide*)”.

The adjacent pairs of these events are {停止-拨付, 拨付-安排, 安排-起诉}, and these relations are shown in the row “Temporal relations of adjacent event pairs”. However, the relation of the adjacent event pair “安排-起诉” is not useful information for readers because the event “安排 (*prepare*)” is a subordinate event of the event “拨付 (*provide*)” and it

describes a past event as a supplement of the event “拨付 (provide).” The temporal relation between events “停止 (stop)” and “起诉 (prosecute)” is more useful than the relation between events “安排 (prepare)” and “起诉 (prosecute)” because events “停止 (stop)” and “起诉 (prosecute)” are coordinate events.



Relation Types	Examples
Temporal relations of adjacent event pairs	{安排, 拨付, after}, {起诉, 安排, after}, {拨付, 停止, before}
Temporal relations of head-modifier event pair	{安排, 拨付, after}, {拨付, 停止, before}
Temporal relations of sibling event pair	{停止, 起诉, simultaneous}
Extend event relations using inference rules	{停止, 安排, after}, {拨付, 起诉, before}
True temporal relations	{安排, 拨付, before}, {安排, 停止, before}, {安排, 起诉, before}, {拨付, 停止, before}, {拨付, 起诉, before}, {停止, 起诉, simultaneous}

Figure 2. The temporal relations in the example phrase

In the example in Figure 2, a native annotator can recognize that the temporal relation between “安排 (*prepare*)” and “起诉 (*prosecute*)” is “before”. However, many event pairs like this example do not have an explicit temporal relation. To analyze this kind of event pairs (“安排 (*prepare*)” and “起诉 (*prosecute*)”), we should consider not only the adjacent observation of events but also dependency structure of sentences to acquire the correct temporal information. Moreover, if an adjacent event pair does not have understandable relation, the adjacent chain (adjacent links) will be segmented. The dependency structure can be used to connect the fragments of adjacent links.

The row “Temporal relations of Head-modifier event pairs” in the table shows the temporal relations of the head-modifier event pairs. We can determine these head-modifier event pairs as subordinate relations. For the event “起诉 (*prosecute*)”, the most important information is the relation between the coordinate event pair “停止 (*stop*)” and “起诉 (*prosecute*)”. We define the event pairs that share a head event as a sibling event pair and show them in the row “Temporal relations of sibling event pairs” of the table. It should be noted that some adjacent event pairs are also head-modifier event pairs or sibling event pairs. The event pairs {拨付-停止, 安排-拨付} are adjacent event pairs and are head-modifier event pairs. Naturally, the event pair should have a similar temporal relationship from a different viewpoint.

### 3.3 Use of Inference Rules

After annotating these relation tags, we use inference rules (See Table 6), such as: “if event A occurs before event B, and event B occurs before event C, then event A occurs before event C”, to extend the temporal relations. The row “Extend event relations using inference rules” shows the temporal relations that are extended using inference rules. By annotating the three types of temporal relations and using the inference rules to extend the temporal relations, we do not need to annotate all possible event pairs, but we can acquire a number of useful temporal relations.

In Sections 2.3 and 2.4, we presented the concept that most of the temporal relations between events are among these three types in English. We expect that these three types of links (Adjacent event pairs, Head-modifier event pair, and Sibling event pair) in Chinese are more important than other links. In the next section, we describe our temporal information annotation guideline for Chinese. Section 5 shows the distribution of tags in our corpus and the coverage of the links induced by dependency structure.

## 4. The Annotation Guideline

This section describes our corpus annotation guideline. First, we introduce basic data and annotation tools. Second, we describe the definition of two temporal entities (events and



signals). Third, we describe the concept of an event tree in an article with dependency structure. Fourth, we describe the attributes of the temporal entities and the temporal relation links. Finally, we compare our criteria and that of TimeML.

#### 4.1 Basic Data and Annotation Tools

To recognize subordinate event pairs and head-modifier event pairs, we needed a dependency-parsed corpus. We used the Penn Chinese Treebank [Palmer *et al.* 2005] as the original data. Since the Penn Chinese Treebank does not include the head-modifier relations, we transformed phrase structures into dependency structures using head rules [Cheng 2005]. The head rules decide the head word of each phrase in the phrase structure, and then the phrase structure becomes a dependency tree. We annotate the temporal attributes and the temporal relations of events on a part of the Penn Chinese Treebank. Our corpus contains 151 Chinese news articles with 7,239 events and 49,691 tokens.

The punctuation “,” usually can be used in the semantic ending of a sentence in Chinese. To distinguish the meaning of the punctuation mark “,” is difficult. We define the end mark of a sentence as the punctuation “。” (a period) in our corpus. The average length of sentences in the Penn Chinese Treebank is 27 words (507,222 words / 18,782 sentences) because a sentence in the Treebank could include several clauses which denote independent events.

We introduce the XML format for our data like TimeBank. We use an XML editor<sup>7</sup> for annotating work. Figure 3 shows the window of the XML editor when we adopt it in our

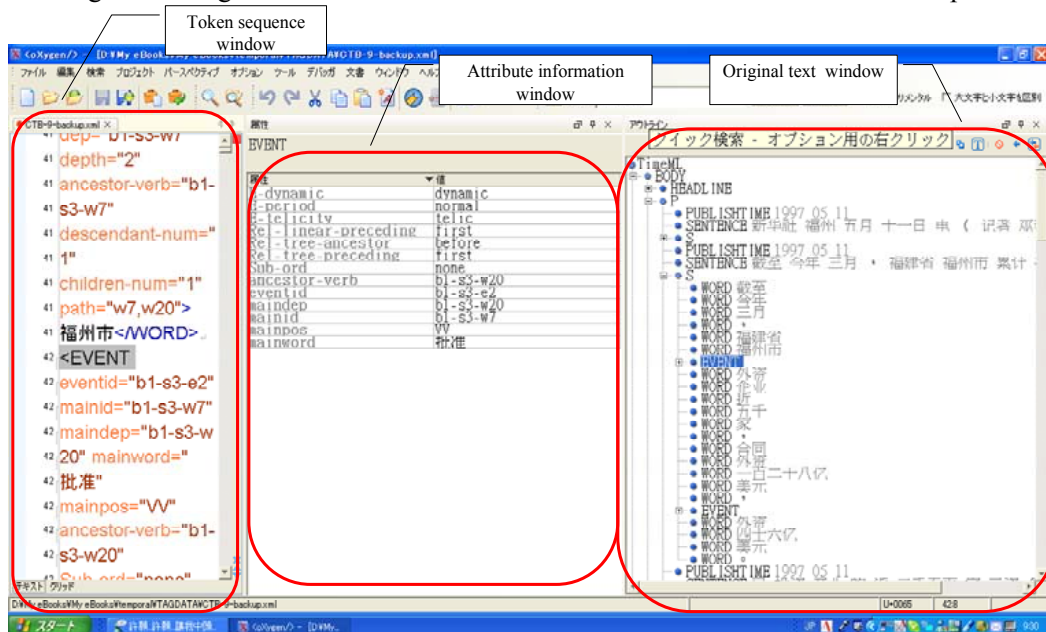


Figure 3. The working window for the annotator

<sup>7</sup> We use the XML editor “<oxygen/>” for our work. (<http://www.oxygenxml.com/>)

annotating work. It includes three sub-windows: “Token sequences window”, “Attribute information window” and “Original text window”. The annotators refer the information in these windows and annotate the attributes of Attribute information windows.

## 4.2 Definition of Events and Signals

We annotate two sorts of entities; one is EVENTS, the other is SIGNALS. The definition of EVENT is based on the TimeML: an entity that describes a situation of happen, occur, state, or circumstance. However, we limit an event to one expressed by a verb in our guideline. According to the guidelines of Treebank, verbs serve as the predicate of a main clause or an embedded clause in corpus [Xia 2000]. We assume that a verb in a clause can be thought of as the representative entity of an event that the clause describes.

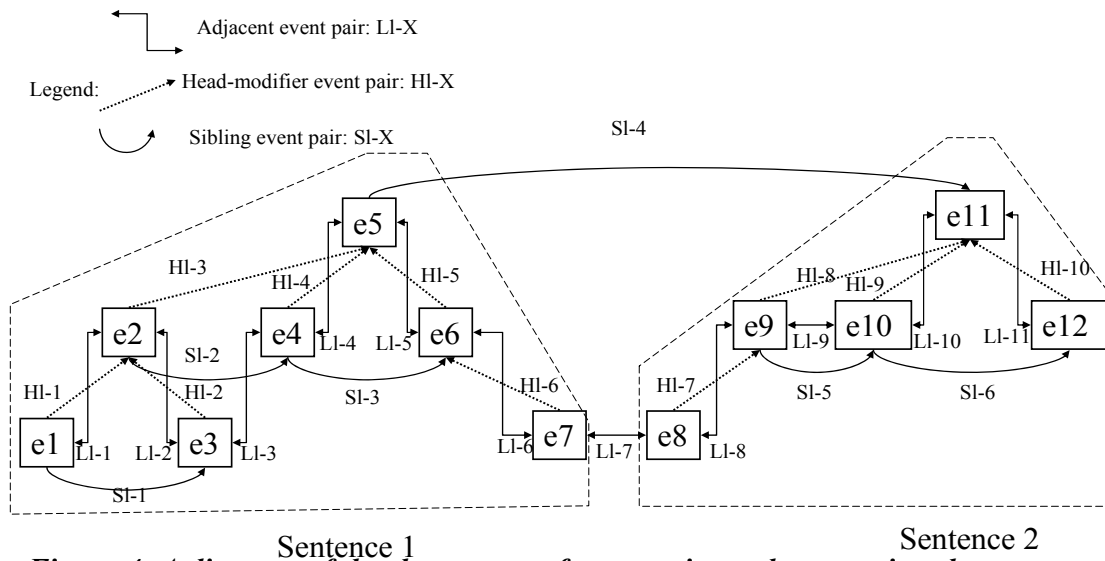
Verbs can be identified automatically, according to the POS tag of the word (the POS-tag: VV, VA, VC and VE). Most of the verbs in Treebank have the POS-tag “VV”, which includes major verbs, such as raising predicates (“可能”(may be)), control verbs (“要”(want)), physical action (“飞”(fly)), psychological action (“讨厌”(hate)), and so on [Xia 2000]. The POS-tag “VA” is used for predicative adjectives, such as “齐全”(well-appointed). We consider predicative adjectives to be the same as an event because these predicative adjectives usually describe a statement. The predicative adjectives can modify a noun in another context, but in these cases the POS-tag of the predicative adjectives is an adjective “JJ”. The difference can be analyzed in the step of morphological analysis or be distinguished in the original Treebank. Therefore, we also recognize predicative adjectives as a type of event according to the POS-tag. The POS-tag “VC” is the copula verb such as “是”(is). It describes a statement of a truth, such as the verb “是”(is) in the sentence: “我是学生”(I am a student), and we define these verbs as EVENT. The POS-tag “VE” describes the possessive or existential statement, such as the verb “有”(have) in the sentence “我有一本书”(I have a book). All these types of verbs are EVENTS and have the annotatable attribute in our criteria.

A SIGNAL is a textual element that makes explicit the relation between two temporal entities. In TimeML, it includes temporal prepositions, temporal conjunctions and prepositions signaling modality. Briefly, the original signals are composed by prepositions or conjunctions. A signal word could mean temporal or non-temporal relations depending on the contextual information. In the sentence “我昨天早上六点起床, 吃过早餐后搭公车上学”(I woke up at 6 A.M. yesterday, ate breakfast and then went to school by bus.) in Figure 1, the word “后”(after, then) is a signal word and describes the fact that the event “吃早餐”(eat breakfast) occurs before the event “搭公车”(by bus). However, the same word “后”(after, then) in the sentence “屋后有个花园”(There is a garden behind the house) means a location relation. Candidate words of SIGNAL in Chinese are limited. We collect these signal candidates according to the POS-tag standard of CKIP’s corpus [CKIP 1993], which lists the SIGNAL

candidates in the POS-tag “Ng” (Localizer). For example, the words “前, 后, 以上, 之前” are signal candidates. However, these signal candidates in Penn Chinese Treebank are not listed and are spread out in the prepositions (“P”), conjunctions (“CC”) and localizers (“LC”). To recognize the SIGNAL automatically, we use the SIGNAL candidate list that is collected from CKIP Treebank to annotate the words that correspond to the list as the “possible signal”. A signal candidate word has an attribute with two classes: “time” and “non-time” to describe if it is a temporal signal word or not. We require the annotator to classify the signal candidate words manually. We will distinguish the use of signal words in different contexts using a machine learning classifier in future research.

### 4.3 A Diagram of the Three Types of Event Pairs

For annotating an article, we transform the parsed sentences into dependency structures. Figure 4 describes the relation of three types of event pairs in an article. There are two sentences with twelve events (from e1 to e12) in the figure, and the polygons with



**Figure 4. A diagram of the three types of event pairs and connecting the sentences**

dashed-lines show the boundaries of sentences. The broken-line links show the adjacent event pairs (from LI-1 to LI-11). The dotted-line links show the head-modifier event pairs (from HI-1 to HI-10), and the curved links show the sibling event pairs (from SI-1 to SI-6). Some adjacent event pairs overlap head-modifier event pairs or sibling event pairs. Most of the three types of temporal relation links are in local structures (in sentence). To connect the temporal relations between sentences, the adjacent event pair links and the sibling event pair links can be used. The link SI-4 and the link LI-7 indicate the links that connect two sentences. The link “SI-4” indicate the relations between the event “e5” and the event “e11”. These events are the matrix events (main events) of “sentence 1” and “sentence 2”. If we postulate that the article

have a dummy root event and this dummy event is the parent event of all matrix events, the relation between the event “e5” and the event “e11” is a sibling event pair. We can use the

属性	
WORD	
属性	值
TMP	
ancestor-verb	b1-s3-w7
children-num	1
dep	b1-s3-w7
depth	2
descendant-num	1
path	w7, w20
pos	NR
signal	
verb-class	
wid	b1-s3-w5
word	福州市

属性	
EVENT	
属性	值
E-dynamic	dynamic
E-period	normal
E-telicity	telic
Rel-linear-pr...	
Rel-tree-ance...	other
Rel-tree-prec...	sub-ord
Sub-ord	explanation
ancestor-verb	b2-s9-w35
eventid	b2-s9-e3
maindep	b2-s9-w15
mainid	b2-s9-w11
mainpos	VV
mainword	鼓励

Figure 5. Attribute windows for annotators.

inference rules on the connecting relations (SI-4 and LI-7) to deduce the temporal relations that cross into adjacent sentences.

#### 4.4 The Attributes of the Entities

We annotate the two types of temporal attributes of events: the properties (dynamic, period and telicity) and the relations for limited event pairs (adjacent event pairs, head-modifier event pairs, and sibling event pairs). Some information of words and events can be annotated automatically, such as the POS-tag, head word, the path to the root of the sentence, and so on. The annotator refers to the annotated information to decide the most appropriate attributes of the temporal lines and event properties of each event. Figure 5 shows the attribute windows.

The left side window in Figure 5 shows the morphological information and the dependency information of a word, and the definition of these attributes is described in Table 3. All of these attributes are analyzed automatically. The “word” tag and the “POS” tag are similar to the original Treebank. The “TMP” tag refers to the phrase tag “\*-TMP<sup>8</sup>” in the Treebank.

The “verb-class” is a concept class of verbs. The verbs in Penn Chinese Treebank include four POS-tags (VV, VA, VC and VE). To give more semantic information of verbs, we define four classes of verbs to describe dynamic concepts: “state”, “change”, “action” and “mental.” The class “state” describes a statement or a static situation, such as “齐全”(well-appointed). Most of the verbs of this class are the verbs with the POS-tag “VA”. The class “change”

<sup>8</sup> The Treebank includes “NP-TMP” (nominal phrase), “PP-TMP” (prepositional phrase), “LCP-TMP” (Localization phrase) and others.

describes the change of statement, such as “变成”(become). The class “mental” describes a psychological action or state, such as “认为”(think) and “讨厌”(hate). The last verb class is “action”. We manually classify 23,979 verbs [GOH 2006] into these four classes. We assume that the concept of verb is important information for recognizing temporal relations. The verbs “射击”(shoot) and “认为”(assume) are in different classes. The class of the former verb “射击”(shoot) is “action” and it usually means a time-bounded action (short period or instantaneous). The class of the latter verb “认为”(assume) is “mental” and it usually means a mental statement with long continuance. It should be noted that the “verb-class” is not necessarily the event class. The property of event could change with context. Actually, we require a lexicon of event semantics, such as Lexical Conceptual Structure [Jackendoff 1992], to classify the verbs in our dictionary. However, there is no Chinese lexicon with event semantic information that covers the verbs in our dictionary. Therefore, we classify the verbs into the four classes manually before we annotate the corpus. The temporal relation annotators are not required to classify the verbs when they annotate the corpus.

**Table 3. Attributes of a word**

Attribute	Definition
The dependency information	
ancestor-verb	The ancestor verb of the focus word
children-num	The number of children of the focus word
Dep	The head word ID of the focus word
Depth	The depth of the focus word in the dependency tree
descendant	The number of descendant of the focus word
path	The path from the focus word to the root of the dependency tree
The morphological information	
TMP	Is the focus word a part of a temporal expression? (yes or no)
pos	POS tag
signal	Is the focus word a signal word? (yes or no)
verb-class	The temporal meaning class of the verb (“state”, “change”, “action” and “mental”)
wid	The ID of the focus word
word	The focus word

The right side window in Figure 5 shows the attributes of the focus event. Table 4 describes the attributes of an event. The attributes of an event include three parts: the information of the main verb in this event, properties of the event (E-dynamic, E-period, and E-telicity), and the temporal relations (Rel-liner-preceding, Rel-tree-preceding, Rel-tree-ancestor and Sub-ord). Properties of an event are the temporal characteristics of the

event; these are different from the concept class of verbs that have already been described. These characteristics roughly correspond to the classification of verbs in Dorr and Olsen [1997]. These properties can describe the verb classification by Vendler [1967] or other classification through the combination of binary values. It includes telicity, dynamic characteristic, and occurrence period of a verb. The value of the tag “E-period” has a special value “forever” and the value “forever” describes an eternal action or situation. For example, the verb “绕” (*circles*) in the sentence “地球绕太阳” (*the earth circles the sun*) has the value “forever”. Although some thesauri contain part of these properties of verbs, they are not publicly available. Additionally, as the properties could change in different contexts; we cannot annotate the properties automatically. We asked annotators not to classify events into several verb classes directly but instead select three binary attributes (E-dynamic, E-period, and E-telicity).

**Table 4. Attributes of an event**

Attribute	values	Definition
information of the main verb		
ancestor-verb		The ancestor verb of the main verb of the event
eventid		The ID of the event
maindep		The head word ID of the focus word
mainid		The ID of the main verb
mainpos		The POS tag of the main verb
mainword		The main verb
the temporal properties of the event		
E-dynamic	state, dynamic	Activity of event
E-period	durative, instantaneous, forever	Period of event
E-telicity	telic, non-telic	Telicity of event
the temporal relation tag of the event		
Rel-linear-preceding	Relations in Figure 6	Relation between the focus event and the linear adjacent preceding event
Rel-tree-preceding	Relations in Figure 6	Relation between the focus event and the sibling event
Rel-tree-ancestor	Relations in Figure 6	Relation between the focus event and the ancestor event
Sub-ord	modal, explanation, condition, none, report	Subordinate type between the focus event and the ancestor event

The four attributes (Rel-linear-preceding, Rel-tree-preceding, Rel-tree-ancestor, and Sub-ord) are temporal relations with another event. We describe these attributes in the following section.


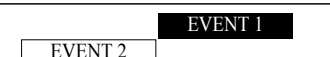











Relation types	Our criterion	TimeML	Allen
	AFTER	AFTER	after
		IAFTER	met-by
	OVERLAPPED-BY		overlapped-by
		ENDS	finishes
	DURING	DURING/IS_INCLUDED	during
	BEGUN_BY	BEGUN_BY	started-by
	SIMULTANEOUS	SIMULTANEOUS/IDENTITY	equal
	INCLUDES	INCLUDES/DURING_INV	contains
	ENDED_BY	ENDED_BY	finished-by
	OVERLAPS		overlaps
		BEGINS	starts
	BEFORE	IBEFORE	meets
		BEFORE	before
Non-temporal relation	first, ambiguous, none	/	

Figure 6. Relation definitions among our criteria, TimeML and Allen's work

#### 4.5 Annotating Links

Our definition of temporal relations is based on TimeML language and Allen's research [Allen 1983]. The original definition of Allen's temporal relations is the relations between two time-intervals. We define four types of temporal relations between two events -- Rel-linear-preceding, Rel-tree-preceding, Rel-tree-ancestor, and Sub-ord. The first three

relations correspond to the relations that we described in Section 3 (“Rel-linear-preceding” refers to the adjacent event pairs, “Rel-tree-ancestor” refers to the head-modifier event pairs and “Rel-tree-preceding” refers to the sibling event pairs). The possible temporal relations are shown in Figure 6. EVENT 1 is the focal event and EVENT 2 is the related event. We group the temporal relation “overlapped-by” and “finished” in Allen’s definition into the temporal relation “OVERLAPPED-BY” in our criteria because there are few instances of “overlapped-by” in our experience. We also group the temporal relation “overlaps” and “start” into the relation “OVERLAP”. The group “AFTER (BEFORE)” includes “after” and “met-by” (“before” and “meet”) in Allen’s definition. This is because distinguishing “after” and “met-by” (“before” and “meet”) is difficult. Except for these groups, other relations are similar to TimeML and Allen’s definitions. In addition to the recognizable temporal relations, we define three un-recognizable classes of relations: “first”, “ambiguous”, and “none”. The value “first” means that the focus event does not have a comparable event. For example, the first event of a sentence does not have a preceding adjacent event; therefore, the value of the attribute “Rel-linear-preceding” is “first”. The value “ambiguous” means that the temporal relation is suitable for more than one relation class. For example, if the event pairs could be annotated as “ENDED\_BY” or “INCLUDES”, the annotator should select the value “ambiguous”. If the temporal relations cannot be decided (such as in an assumptive situation), the annotator is asked to select the value “none” for the focus event.

**Table 5. Definition of subordinate class**

subordinate class	definition
Modal	The focus event is introduced by the head event.
explanation	The focus event explains the head event.
condition	The focus event occurs if the head event is true.
report	The head event is a “report” event.
passive	The focus event is passive of the head event.
possibility	The head event describes a possibility of focus event.

The last relation “Sub-ord” means the subordinate relation of a head-modifier event pair. We refer to TimeML in defining the subordinate relations. The annotator should annotate this relation without depending on the three temporal relations – “Rel-linear-preceding”, “Rel-tree-preceding” and “Rel-tree-ancestor”. Annotators can refer to the dependency structure of the focus event to recognize the subordinate event and its main event. The definition of the subordinate relations is described in Table 5. TimeML includes another link tag “ALINK” to annotate aspectual relations. We do not distinguish SLINK and ALINK and designate these two kinds of relations as the tag “Sub-ord”. We assume that any subordinate relation could include a temporal relation. As the temporal relations include aspectual



relations (such as BEGUN\_BY and END\_BY), the annotators can annotate the temporal relation between a sub-ordinate event and its head event to cover SLINK and ALINK.

**Table 6. Inference rules**

	The relation between event B and event C				
The relation between event A and event B	AFTER	BEFORE	DURING	INCLUDE	SIMULTANEOUS
AFTER	AFTER			AFTER	AFTER
BEFORE		BEFORE		BEFORE	BEFORE
DURING	AFTER	BEFORE	DURING		DURING
INCLUDE				INCLUDE	INCLUDE
SIMULTANEOUS	AFTER	BEFORE	DURING	INCLUDE	SIMULTANEOUS
	The relation between event A and event C				

**Table 7. The attributes of the events in Figure 2**

event \ Attribute	停止 (stop)	拨付 (provide)	安排 (prepare)	起诉 (prosecute)
the temporal properties of the event				
E-dynamic	dynamic	Dynamic	dynamic	dynamic
E-period	instantaneous	Durative	instantaneous	instantaneous
E-telicity	telic	Telic	telic	telic
the temporal relation tag of the event				
Rel-linear-preceding	first	END_BY	BEFORE	AFTER
Rel-tree-preceding	first	First	first	SIMULTANEOUS
Rel-tree-ancestor	first	END_BY	BEFORE	first
Sub-ord	none	Explanation	explanation	none

After we annotate the aforementioned temporal relations, we can use the inference rules to extend to more temporal relations. Table 6 shows the inference rules that we use in our experiment in Section 5.2. For example, if two temporal relations “Event A occurs during Event B” and “Event B occurs before Event C” are extracted, we can infer a new relation “Event A occurs before Event C”. We will describe the inference rules in more detail in Section 5.2.

Table 7 describes the attributes of the events in the example sentence of Figure 2. In this example, the events “停止” (*stop*) and “起诉” (*prosecute*) are coordinate events, therefore the attribute “Rel-tree-preceding” of the event “起诉” (*prosecute*) is “SIMULATANEOUS”. The attributes “Rel-tree-preceding” of the other events are “first”. The ancestor event of the events

“拨付”(provide) and “安排”(prepare) is the same as their linear adjacent events, therefore the value of the tag “Rel-linear-preceding” is the same as the tag “Rel-tree-ancestor”. Since the event “停止”(stop) is the first event and is the root event of the dependency tree, it does not have linear adjacent events nor ancestor events. Therefore, the values of tag “Rel-linear-preceding” and “Rel-tree-ancestor” are “first”.

#### 4.6 Difference between our Guideline and TimeML

Our corpus guidelines adopt many concepts and attribute values token from TimeML. However, our corpus criteria have several features that are different from TimeML. First, the goal of our research is to construct a machine learning based annotation system. All attributes can be annotated automatically after we complete a large corpus and train a machine learner. In TimeML, the annotators need to extract the temporal entities and relations using their knowledge, which requires a large amount of time. However, in our criteria, our annotators focus on the attributes of events and inference rules are used to extend temporal relations. We can create a large corpus according to our criteria, which would be more difficult using TimeML. Second, for recognizing the events of corpus automatically, we limit the events to the verbs, but TimeML includes more syntactic event constituents. To limit the event only to verbs can reduce the manual effort and preserve the major parts of all events.

When we use the temporal relation tagged corpus to train a machine learner, every attribute of our criteria can be trained. However, training the machine learner with a corpus tagged by the TimeML annotation scheme is more difficult than with our corpus. TimeML includes more difficult criteria. For example, the machine learner should identify the event phrase<sup>9</sup> (clause) in corpus. In a shared task of SemEval-2007 (Task 15: TempEval Temporal Relation Identification) [Verhagen 2007], participants use TimeBank as the corpus to identify the temporal relations in articles. However, they are not required to identify event phrases. Identifying event phrases requires chunking technology and world knowledge.

Finally, our criteria are based on dependency structure. TimeML does not consider any syntactic nor morphological information in their annotation. Our criteria are based on dependency structure and can describe the temporal relations of subordinate and coordinate event pairs clearly. In our experience, these criteria can provide annotators with useful information that help them to recognize the relations between events. Moreover, because verbs in Chinese do not have morphological change according to tense, recognizing the tense of an event needs the information of the modifier of the verb, such as a temporal expression “昨天”(yesterday) or a temporal function word “已經”(have). This information directs the temporal relations analysis for Chinese. This information can be provided by dependency

---

<sup>9</sup> The events in TimeML could be a noun phrase or verb phrase.

structure, whereas TimeML does not emphasize this. Therefore, our criteria are more applicable than TimeML to the creation of a temporal relation tagged corpus of Chinese.

## 5. The Corpus Distribution

In this section, we report the distribution of the corpus annotation. We annotated a part of Penn Chinese Treebank and investigated the distribution of each attribute. Next, we investigated the coverage of the temporal relations using our proposed guideline.

### 5.1 Distribution of Attributes in Tagged Corpus

The Penn Chinese Treebank 5.0 contains 507,222 tokens, 18,782 sentences, and 890 articles. We will automatically analyze these attributes in the future. However, we need manually tagged training data to construct machine learning models. We use a part of the Penn Chinese Treebank (about 10%) to construct a basic data set. As the consistency of the annotated corpus is not competent, we could not use it to get machine learning models before we repeated the annotating work to improve the consistency.

**Table 8. Distribution of the attributes**

Attribute	Values	Number
E-dynamic	State / dynamic	5347 / 1892
E-period	durative / instantaneous / forever	3024 / 4156 / 59
E-telicity	telic / non-telic	3440 / 3799
Rel-linear-preceding	(top four relations) after / simultaneous / before / during	2523 / 2065 / 1091 / 463
Rel-tree-preceding	(top four relations) first / after / simultaneous / before	5116 / 818 / 491 / 305
Rel-tree-ancestor	(top four relations) first / simultaneous / before / after	1968 / 1816 / 1773 / 1073
Sub-ord	(top four relations) none / explanations / modal / report	3622 / 1861 / 556 / 432

The distribution of the attributes is summarized in Table 8. As the distribution of temporal relations is uneven and the space of this paper limited, we only show the top four types of temporal relations. Considering the tag “Rel-linear-preceding (adjacent event pairs)”, the relation classes “AFTER / SIMULATANEOUS / BEFORE” are the most possible relations among the adjacent event pairs. Since we request the annotators to annotate as many temporal relations as possible, they used a considerable amount of world knowledge and contextual information in reading the articles. Therefore, the class “ambiguous” in tag “Rel-linear-preceding” is infrequent. The relation class “first” of the tag “Rel-tree-preceding (sibling event pairs)” means the focus event does not have any sibling event because events in

similar sentences are structured in a hierarchical structure. There are few sentences that have events that modify the same event. Therefore, most events are singletons of their head events. In the tag “Rel-tree-ancestor (head-modifier event pairs),” the root event of the dependency structure does not have a head event and the correct selection of the tag “Rel-tree-ancestor” should be “first” in this case. In the tag “sub-ord (subordinate relation),” the value of the most meaningful types of subordinate relation are “explanations” and the majority of this attribute consists of the tag “none”.

## 5.2 The Coverage of the Links

We investigate a small corpus to observe the performance of our criteria by comparing the results of our criteria and all possible event pairs. As described in Section 3, for  $n$  events in an article,  $nC_2$  relations should be considered<sup>10</sup>. We can compare the relations of all pairs of events and the relations extracted by our criteria to observe the coverage of our criteria. However, it is difficult to annotate the temporal relations of all event pairs. For example, if an article contains 50 events, there are 1,225 event pairs ( $50C_2$ ). We cannot compare the two methods in a large corpus because the annotation cost is huge. Therefore, we select 50 articles in Penn Chinese Treebank and only use the first two paragraphs of each article to make our survey data. The small corpus includes 732 events (verbs) and 5,010 tokens.

We annotate the small corpus manually both by annotating all event pairs and by using our criteria. After annotation by our criteria, we use the inference rules shown in Table 6 to extend the relations. In previous research [Mani *et al.* 2006b], the inference rules could adopt some syntactic or semantic features<sup>11</sup> of event pairs to extend more inference rules. To use syntactic/semantic features, experimental linguistic knowledge is needed to make an induction and we have not collected the linguistic knowledge yet. In this paper, therefore, we use the inference rules that only adopt unambiguous relations without syntactic/semantic features.

To observe the coverage of different methods, we survey four methods of extracting temporal relations. They are: 1. Using the relationships of the adjacent event pairs (RLP is an abbreviation of Rel-linear-preceding), the head-modifier event pairs (RTA is an abbreviation

---

<sup>10</sup> We assume that the inverted relation pairs, such as “event A occurs before event B” and “event B occurs after event A”, are different, because the combination  $nC_2$  only calculates a single direction of temporal relations, for example, the relation that forms event A to event B is extracted, but the relation that from event B to Event A is not considered. However, our method would extract two directions of temporal relation (the relation that from Event C to Event A is possible). If a relation between a combination event pair is extracted, we extend the inverse relation automatically (if event A occurs before event B, then event B occurs after event A).

<sup>11</sup> Such as the “POS” tag and the “TENSE” tag are used for creating inference rules in [McDonald *et al.* 2005].

of Rel-tree-ancestor), and the sibling event pairs (RTP is an abbreviation of Rel-tree-preceding), then extending the relations by the inference rules (The column “RLP+RTA+RTP” in Table 9). 2. Only using the relations of the adjacent event pairs with the inference rules (The column “RLP” in Table 9). 3. Using the relations of the head-modifier event pairs and the sibling event pairs with the inference rules (The column “RTA+RTP” in Table 9). 4. Using three kinds of event pairs without the inference rules (The column “RLP+RTA+RTP w/o inference rules” in Table 9). For experimental convenience, we reduce the nine classes of temporal relations to five classes. The classes {AFTER, OVERLAP\_BY, BEGUN\_BY} are reduced to the class “AFTER” and the classes {BEFORE, OVERLAP, ENDED\_BY} are reduced to the class “BEFORE.” According to our annotator’s experience, these subclasses are ambiguous in many event pairs; therefore, we group the classes to reduce the ambiguity.

**Table 9. Results of the coverage evaluation**

	RLP+RTA+RTP	RLP	RTA+RTP	RLP+RTA+RTP w/o inference rules
Relations of Adjacent event pair (The attribute Rel-linear-preveding-RLP)	702	702	0	702
Relations of Head-modifier event pair (The attribute Rel-tree-ancestor-RTA)	530	0	530	530
Relations of Sibling event pair (The attribute Rel-tree-preveding-RTP)	205	0	205	205
Total extracted event relations	1018	702	735	1018
Extend event relations by using inference rules	4166	2005	2871	1018
True event pairs	6646	6646	6646	6646
Recall	0.63	0.30	0.43	0.15

Table 9 describes the coverage of our proposed methods. We regard the understandable relations of all event pairs as the gold standard (the row “True event pairs”) and we compare the result of our method with the gold standard. The row “Recall” shows the coverage of each method. We do not show the accuracy of each column because the extended relations are all included in the “True event pairs” and the precision of each column is “100%”.

The last column shows the case of using our criteria to annotate temporal relations without using the inference rules. The row “Extend event pair relations using the inference rules” in this column indicates the total number of events that are annotated by our criteria. It should be noted that an adjacent event pair could be also a sibling event pair or a head-modifier event pair. For example, the event pair the event “安排 (*prepare*)” and the

event “拨付 (*provide*)” is both an adjacent event pair and a head-modifier event pair. It will be calculated twice in the two types of event pairs. Therefore, the number of the relations that we extract by our criteria is not equal to the total number of the three kinds of relation types ( $RLP+RTA+RTP > \text{Total event pairs}$ ).

Intuitively, the combination of events must include all relations that could be extracted. The relations that we extract by our criteria must be included in the gold standard. In Table 9, the row “Total extracted event relations” is included in the true event relation. However, in our preliminary investigation (this result is NOT included in Table 9); the annotator does not consider any syntactic structure or full context in annotating the event pairs and then the extracted event relations are not completely consistent to the true event relations. Because this testing data set was annotated by an annotator but not completed in one day, the annotator does not remember the viewpoint before when he annotates the same instance. The annotator annotated the event combination first and then annotated the three types of event pairs of our criteria. The intuitive reorganization of event relations could be inconsistent with the dependency structure. Therefore we re-annotated the testing data several times to confirm the consistency of the relation attributes. This observation indicates the difficulty of constructing a corpus consistently.

According to our results, the precision of using “RTA+RTP” with the inference rules is better than the precision of only using “RLP” with the inference rules. The hypothesis in Section 3 is confirmed in the result. The head-modifier event pairs can connect some fragment structure and can extract many important relations that the adjacent event pairs cannot extract. The recall row shows the coverage of our method. We use three types of event pairs and the inference rules and acquire 63% relations of the gold standard can be extracted. One reason is that we only consider the absolute inference rules. We can add more inference rules that consider other syntactic or semantic information of events to extend the relations.

## 6. Conclusion and Future Direction

This research focuses on an annotation guidelines for a temporal relation tagged corpus of Chinese. The guidelines are based on the TimeML language, but we also use dependency structure information to acquire more meaningful temporal relations and to reduce manual effort. We define events as those expressed by verbs and define three types of links for event pairs. These types (the adjacent event pairs, the head-modifier event pairs, and the sibling event pairs) include most meaningful information, and we extend these relations using the inference rules.

To annotate temporal relations of all combinations of events requires  $nC2$  manual judges. Our criteria require at most  $3n$  times of annotation. While the dependency structure based attributes reduce manual annotation costs, the limited relations preserve the majority of the

temporal relations. The average working time required for one article (with 80 events) is about 30 minutes in our annotation work. It is shorter than the annotating work of TimeBank [Pustejovsky *et al.* 2006], which is 45 minutes for one article. We survey the coverage of our method with a small corpus. The result shows that our method covers about 63% of temporal relations. We expect that extension by our inference rules enables one to extract more temporal relations.

In future research, we will use the machine learning method to annotate the temporal relations. We annotated the Penn Chinese Treebank ceaselessly by our criteria. Once we annotate enough data, we will train it and, thereby, should reduce the inconsistency of our data.

## References

- Allen, J. F., "Maintaining Knowledge about Temporal Intervals," *Communications of the ACM* (Association for Computing Machinery), 26(11), 1983, pp. 832-843.
- Bach, E., "the algebra of events", *Linguistics and Philosophy* 9, Swets & Zeitlinger Publishers, London, 1986, pp. 5-16.
- Brants, T., TnT - A Statistical Part-of-Speech Tagger, <http://www.coli.uni-saarland.de/~thorsten/tnt/>, 1998.
- Cheng, Y., Chinese Deterministic Dependency Analyzer: Examining Effects of Chunking, Root node finder and Global Features, MD thesis, NAIST, 2005.
- Chinchor, N., MUC-7 named entity task definition, [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/proceedings/muc\\_7\\_proceedings/overview.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html), 1997.
- Chklovski, T., and P. Pantel, "VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations," In *Proceeding of 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, Barcelona, Spain, pp. 33-40.
- CKIP, "中文詞類分析," Technical report no. 93-05, Institute of Information Science Academia Sinica, Teipei, 1993.
- Dorr, B. J., and M. B. Olsen, "Deriving Verbal and Compositional Lexical Aspect for NLP Application," In *Proceeding of 35th Annual Meeting of the Association for Computational Linguistics*, 1997, Madrid, Spain, pp. 151-158.
- GOH, C.-L., Unknown Word Identification for Chinese Morphological Analysis, PhD thesis, NAIST, 2006.
- IREX Committee, Named entity extraction task definition, <http://nlp.cs.nyu.edu/irex/NE/df990214.txt>, 1999.
- Jackendoff, R., *Semantic Structure*, The MIT Press, Cambridge, 1992.

- Li, W., K.-F. Wong, and C. Yuan, "Application and Difficulty of Natural Language Processing in Chinese Temporal Information Extraction," In *Proceeding of the Sixth Natural Language Processing Pacific Rim Symposium*, 2001, Tokyo, Japan, pp. 501-506.
- Li, W., K.-F. Wong, G. Cao, and C. Yuan, "Applying Machine Learning to Chinese Temporal Relation Resolution," In *Proceeding of 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, Barcelona, Spain, pp. 582-588.
- Mani, I., J. Pustejovsky, and R. Gaizauskas (ed.), *the language of time*, Oxford University press, Oxford, 2006a.
- Mani, I., M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky, "Machine Learning of Temporal Relations," In *Proceeding of the joint conference of 21st International Committee on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, Sydney, Australia, pp. 753-760.
- McDonald, R., F. Pereira, K. Ribarov, and J. Hajic, "Non-Projective Dependency Parsing using Spanning Tree Algorithms," In *Proceeding of 2005 Conference on Empirical Methods in Natural Language Processing*, 2005, Vancouver, Canada, pp. 523-530.
- Palmer, M., F.-D. Chiou, N. Xue, and T.-K. Lee, Chinese Treebank, version 5.1, <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T01U01>, 2005.
- Pustejovsky, J., M. Verhagen, R. Sauri, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer, TimeBank, version 1.2, <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08>, 2006.
- Sauri, R., J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky, TimeML Annotation Guidelines, version 1.2.1, <http://www.timeml.org/site/publications/specs.html>, 2005.
- Vendler, Z., *Verbs and Times, Linguistics in Philosophy*, Cornell University Press, Ithaca, 1967.
- Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky, "SemEval-2007 Task 15: TempEval Temporal Relation Identification," In *Proceeding of 4th International Workshop on Semantic Evaluations*, 2007, Prague, Czech, pp. 75-80.
- Wu, M., W. Li, Q. Lu, and B. Li, "CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information," In *Proceeding of 2nd International Joint Conference on Natural Language Processing*, 2005, Jeju Island, Korea, pp. 694-706.
- Xia, F., The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank, <http://www.cis.upenn.edu/~chin>, 2000.



# The Effects of Formal Schema on Reading Comprehension—An Experiment with Chinese EFL Readers

Xiaoyan Zhang\*

## Abstract

This study attempts to explore the effects of formal schemata or rhetorical patterns on reading comprehension through detailed analysis of a case study of 45 non-English majors from X University. The subjects were selected from three classes of comparable English level and were divided into three groups. Each group was asked to recall the text and finish a cloze test after reading one of three versions of a passage with identical content but different formal schemata: description schema, comparison and contrast schema, and problem-solution schema. Both quantitative and qualitative analyses of the recall protocol indicate that subjects displayed better recall of the text with highly structured schema than the one with loosely controlled schema, which suggests that formal schemata has a significant effect on written communication and the teaching of formal schemata to students is necessary to enhance their writing ability.

**Keywords:** Formal Schema, Schema Theory, Reading Comprehension

## 1. Introduction

For many people, reading is the most important of the four skills in a second language, especially in English as a second or foreign language. According to Carrell [2006], effective reading in a second language is critical for students in EFL contexts, at an advanced level of proficiency, or with a need for English for Academic Purposes. For a long time, EFL reading was viewed as a rather passive, bottom-up process. In other words, EFL reading was primarily a decoding process of reconstructing the author's intended meaning through identifying the printed letters and words and building up a meaning for a text from the smallest textual units at the "bottom" (letters and words) to larger and larger units at the "top" (phrases, clauses,

---

\*Xi'an University of Finance and Economics, Xi'an, China

E-mail: zhgydyx@yahoo.com.cn

intersentential linkages). Correspondingly, problems of second language reading and reading comprehension were considered essentially decoding problems, deriving meaning from print [Rivers 1964, 1968; Plaister 1968; Yorio 1971].

Only since 1979, has a truly top-down approach been proposed in second language reading [Steffensen *et al.* 1979; Carrell 1981, 1982; Carrell and Eisterhold 1983; Johnson 1981, 1982; Hudson 1982]. The top-down perspective of reading process has had a profound impact on reading comprehension, and it views the top-down perspective of the reading process as a substitute for the bottom-up, decoding view of the reading process, rather than being its complement. Only after the appearance of the schema theory, has it been made clear that effective and efficient reading – either in a first or second language — requires both top-down and bottom-up strategies operating interactively [Rumelhart 1977a, 1980; Sanford and Garrod 1981; van Dijk and Kintsch 1983].

Rumelhart [1977b] views reading comprehension as the process of choosing and verifying conceptual schemata for the text. A schema is said to be “a cognitive template against which new inputs can be matched and in terms of which they can be comprehended” [Rumelhart and Ortony 1977: 131]. According to the schema theory, not only is the reader’s prior linguistic knowledge (linguistic schemata) and level of proficiency in the second language important, but the reader’s prior background knowledge of the content area of the text (“content” schemata) as well as of the rhetorical structure of the text (“formal” schemata) are also important [Carrell 2006].

The importance of linguistic schemata in reading comprehension has long been recognized because of the long history of the bottom up view of reading comprehension, and, with the rise of the schema theory, researchers have showed great interest in the importance of content schemata and formal schemata. However, compared with the studies on content schema, studies on formal schema are much less frequent. Therefore, the present study focused on the effects of formal schema on reading comprehension with a view to arriving at a better understanding of the unique nature of EFL reading.

This paper consists of five parts. Part One provides an overview of the study. Part Two presents the background of the study. The elaboration of the experiment is found in Part Three, while the results and discussion are presented in Part Four. Part Five summarizes the findings of the experiment and discusses the implications of the results to the development of EFL reading instruction as well as the limitations of this study.

## **2. Literature Review**

Formal schemata [Sharp 2002] are part of the macrostructure of a text and contain the logical organization of the text which the writer has used to represent the intended meaning.

Meyer and Freedle [1979] explored the effects of different formal schemata on recall. The 4 types of formal schemata compared were: (1) contrastive schema; (2) cause-effect schema; (3) problem-solution schema; and (4) collection-of-descriptions schema. The first three types of formal schemata have “an extra link of relationship” over the descriptive schema. Results demonstrated that subjects who were exposed to formal schemata 1 and 2 recalled more than formal schemata 3 and 4. The results can be explained by schema theory. Based on this theory, recall of information relayed by the first three formal schemata, which offer extra linkage, should be better than that of the descriptive schema. Meyer *et al.* [1980] conducted another experiment to confirm that readers who adopted the strategy of identifying the author’s organization structure would be able to recall more information than students who did not. Results were consistent with the predicted outcome.

The experiments described above were conducted on L1 readers. Research dealing with L2 readers has been more limited. However, Carrell [1984] used Meyer’s passage on “Loss of body water” presented in different schemata to test the effects of top-level organization on ESL readers. Results indicated that certain highly structured schemata facilitated the recall for L2 readers in general. L2 readers tested included Spanish, Arabic, Oriental languages (Korean and Chinese), and Malaysian. Carrell’s work was duplicated by Foo [1989], Goh [1990], and Talbot *et al.* [1991] using exactly the same texts. For all of these investigations, the general conclusion was that the formal schemata of the texts had an effect on reading comprehension as measured by recall and the more tightly controlled schema seemed to be easier to comprehend. Perego and Boyle [2000] also stated that text structure knowledge enhanced comprehension by helping readers to anticipate and predict the direction of a plot or argument, thereby facilitating attention to the larger meaning of the text. Besides, Sharp [2002] studied the effects of the four formal schemata of expositions on comprehension by cloze test and recall protocol. The four formal schemata include description, cause-effect, listing and problem-solution. The result was consistent with the previous work in that it clearly demonstrates that formal schemata do affect reading comprehension. However, the results indicated that it was the most loosely organized text (description) that scored the highest, which was quite different from other investigations of this type [*e. g.* Carrell 1984; Foo 1989; Goh 1990].

The above-mentioned studies have provided a general view upon the role that formal schema plays in EFL reading comprehension. However, some aspects still remain unexplored. First, some results were quite contradictory, for example, Sharp [2002] versus Carrell [1984] as well as Foo [1989]. Besides, no study has been conducted to study the effects of formal schemata in English reading on a specific cultural group — Chinese EFL non-English college students.

A text is a complete linguistic unit to discuss a topic, while around the topic different

people have different ways of developing it [Xiao 2001]. Kaplan [1966] claims that the rhetorical pattern of a text is language-specific and culture-unique and reflects the thought pattern of a particular group. For instance, the English thought pattern is straightforward and the topic is usually developed using a deductive method, while the Chinese thought pattern is spiral and the topic is usually developed via an inductive method [Xiao 2001]. Therefore, the question this study addresses is:

Will the formal schemata of EFL (represented by three different patterns) affect the reading comprehension of native Chinese non-English major college students?

### **3. Method**

#### **3.1 Subjects**

Fifty-five sophomores from the School of Liberal Arts and Law of X University were selected to participate in this study. They ranged from 17 to 19 years old, and 15 of them were male. Before entering university, these students already had six years of contact with English as a foreign language, with an average of four hours of English classes per week. By the time they entered university, they had learned about 2,000 words, and they could read simplified English texts and write short compositions. As non-English majors at X University, the sophomores took a College English course totaling 4 hours per week.

The 45 participants were selected out of 150 students from three classes in School of Liberal Arts and Law and students from the same class fell into the same group (each group 15 participants). The researcher excluded the top 35 students and poor students and chose those 15 intermediary students in each class on the basis of the final exam of the English course. In addition, a one-way ANOVA indicated no significant differences existed among three groups of subjects.

#### **3.2 Material**

Three versions of a passage with identical content but different formal schemata were used (see Appendix A). The three types of formal schemata selected in the study are description, comparison and contrast, and problem-solution. This combination is unique for a couple of reasons: the previously-mentioned studies [Sharp 2002; Foo 1989; Carrell 1984] all studied the first and the last type of schemata while bringing out contradictory results, along with the comparison and contrast type being missed in both studies [Sharp 2002; Foo 1989]. Passage 1 is description, and no clear relationship can be seen among its components. Therefore, it is highly loosely organized. The macrostructure of Passage 2 (comparison and contrast) follows the three-part pattern — introduction, body, and conclusion. In the introduction part, not only are the subjects to be compared and contrasted identified, but also the points made about them

are stated clearly. The body adopts the subject-by-subject pattern. One subject is to be discussed fully, then the next subject. In discussing each subject, some aspects are examined in detail. The conclusion part summarizes the main points. Passage 3 is problem-solution which is arranged into three parts. The introduction introduces the problem, the body expands on the solutions to the problem, and the conclusion is the summary of the solutions to the problem. The discourse signals in Passage 2 and Passage 3 are highlighted in Appendix A.

The three passages are appraised as the same level of difficulty by three English professors in Y University in terms of word length, word choice, and sentence type. The numbers of words in the three passages were 192, 193, and 196, respectively. No uncommon words appeared in any of those passages. There were four complex sentences in each of the three passages.

### **3.3 Instruments**

A considerable number of methods have been used to measure reading behavior and reading comprehension. A text-based cloze procedure and a recall protocol were considered the most appropriate methods for this study. Both methods have been used in reading comprehension investigation extensively, and both methods allow a large number of subjects to be assessed.

#### **3.3.1 Cloze Test**

A cloze test was employed in this study. Cloze tests have aroused much discussion and have drawn lots of criticism. However, there is enough evidence to support them as a measure of reading comprehension. Bormouth [1968] concluded that cloze tests seemed to be valid measures of passage difficulty. Cloze tests appear to be valid, reliable language proficiency tests that can be easily constructed and used by ESL teachers [Aitken 1975; Stubbs and Tucker 1974; Oller 1973; Brown 1980]. Besides, both Bachman [1985] and Jonz [1990] reviewed the investigations supporting cloze.

The type of cloze construction adopted in the present study is similar to the one in Sharp's study [2002], based on Farhady and Keramati's study [1996]. Farhady and Keramati's design calculates deletion rates on the basis of noun phrases in a text and they assert that such a design takes better account of the discorsal and linguistic structure of the language used and is a comparatively good test of reading comprehension because of improved reliability and validity. In this design, noun phrase calculation should conform to the following rules: conjoined NPs were regarded as single units; complex NPs (NPs with embedded NPs) were counted as single units and pronouns were ignored. Exact word scoring, which requires the word put in to be the exact word used in the original text, was used. Deletion rates for the text were: *description* every 6<sup>th</sup> word, 29 deletions; *comparison and contrast* every 7<sup>th</sup> word, 25 deletions; *problem-solution* every 6<sup>th</sup> word, 30 deletions. Deletion rate changes between 6 and

7 are not likely to have an effect [Alderson 1979; Porter 1978]. A sample cloze together with calculation is shown in Appendix B.

### 3.3.2 Recall Protocol

This study also used a recall protocol to measure subjects' reading comprehension of the given texts. Employing a recall protocol to test reading comprehension is a common practice in research of this sort. For example, Morrow [1988] and Salvia and Hughes [1990] strongly recommended recall as a method of classroom assessment for instructional and diagnostic purposes. This method requires that the text be divided into idea units. An idea unit, also called a linguistic unit [Bransford and Franks 1971; Carrell 1983] and an information unit [Roller 1990], is the minimal words necessary to express a thought or idea. Therefore, subjects' reading comprehension is measured by the number of idea units recalled, *i.e.* the amount of information recalled. The segmentation of texts in this study was similar to Johnson's [1970] to allow quantitative assessment of recall. Furthermore, to account for qualitative level differences in recall, the researcher followed Sharp's practice [2002]: the idea units were also rated for importance within the text. The quantity and quality of idea units was determined by the agreement reached by eight English professors in Y University and was shown in Table 1 (see Appendix C).

### 3.4 Data Collection

At the very beginning of the test, subjects were briefly informed that the text was about houses. Since recall protocols allow the possibility of rote learning without real understanding of the text, they were not told that they would be asked to recall and they were asked to fill a questionnaire about personal information (see Appendix D) as a distraction task between the initial reading of the text and the recalling of the text. When the test started, the three texts with different formal schemata were distributed to the three groups of subjects, respectively, and the first sentences of all three texts were the same. Then, they were asked to read the text in five minutes. After three minutes of answering the questionnaire, they were told that that they should take a test in English within ten minutes. Eventually, a ten-minute cloze test was administered.

### 3.5 Data Analysis

The 45 cloze items (15 descriptions, 15 comparison and contrasts, and 15 problem-solutions) were rated by two raters. They were completely agreeable with each other because cloze requires exact words. In addition, the 45 recall protocols were also sent to these two raters to assess both in terms of quantity and quality on the basis of the templates reached by eight English professors in Y University. For the convenience of comparison, both the quantitative

score and qualitative score of recall protocols adopted the percentage system. The quantitative score of each recall protocol was calculated according to the following formula: quantitative score = (the idea units recalled / the total idea units in the recalled passage)×100. The qualitative score of each recall protocol were obtained based on another formula: quantitative score = (sum of the importance level of each recalled unit / sum of the importance level of all idea units in the recalled passage)×100. In this study, misspellings and grammatical mistakes did not affect the score of a subject's recall because such mistakes did not mirror readers' understanding of the passage. The reliability of the essay ratings for the two raters was excellent (see Table 2) and the final score of each one was the average of the two scores given by the two raters. Both raters selected for this study were intensive reading teachers of English with more than five years in the English department of Y University, so they were highly competent to fulfill this task.

**Table 2. Interrater Reliability for the Two Raters**

Reading text	Correlation coefficient 1 (quantity)	Correlation coefficient 2 (quality)
Text 1	0.97	0.94
Text 2	0.88	0.87
Text 3	0.93	0.89

The quantitative and qualitative scores of recall protocols were input in Statistical Package for Social Sciences (SPSS) for descriptive and inferential analysis. The effect of formal schema on reading comprehension was measured through one way ANOVA.

#### 4. Results and Discussion

Research Question:

Will the rhetorical pattern of EFL (represented by three different patterns) affect the reading comprehension of native Chinese non-English major college students?

**Table 3. Cloze, Recall 1 (Quantitative) and Recall 2 (Qualitative) scores for the 3 texts**

Text	Cloze Mean	Recall 1 Mean (Quantitative)	Recall 2 Mean (Qualitative)
Text 1	51.0000	39.6667	43.5333
Text 2	53.3333	48.2667	49.9333
Text 3	59.5333	58.4667	62.6667
Overall	54.6222	48.8000	52.0444
F	1.120	8.863	8.373
Sig.	.336	.001	.001

The means of three kinds of score: cloze, recall 1 (quantitative), and recall 2 (qualitative) for the three texts with different formal schemata (description, comparison and contrast, problem-solution) are presented in Table 3. As shown in Table 3, the cloze test did not demonstrate significant differences among the three schematically different texts ( $p=0.336>0.05$ ). However, when it comes to recall protocol, both the quantitative and qualitative scores indicated significant differences among the text types. In terms of quantitative measure the three schematically texts were significantly different at the level of 0.001, with Text 1 (description) scoring the lowest (mean=39.6667) and Text 3 scoring the highest (mean=58.4667). Text 2 scored in between (mean=48.2667). As for qualitative measure, the texts again witnessed significant differences at the level of 0.001. Text 3 (mean=62.6667) still obtained the highest score while Text 1 (mean=43.5333) achieved the lowest and Text 2 was in the middle (mean=49.9333).

This means that, in terms of the cloze test, no significant difference could be found among the three texts with different schemata, but, according to the number and the importance of the idea units recalled, significant differences did exist among them.

In general, the effect of formal schemata on reading comprehension observed in this study is quite similar to the profile documented in many research papers [Meyer and Freedle 1979; Meyer *et al.* 1980; Carrell 1984; Foo 1989; Goh 1990; Talbot *et al.* 1991; Sharp 2002] in that it revealed that formal schemata indeed affects reading comprehension. However, in this study, no significant difference was found in the cloze test among the schematically different texts as opposed to Sharp's study [2002], in which the four text types — description, cause-effect, listing, and problem-solution differed significantly in the cloze score with description scoring the highest. This can be attributed to the teaching and learning style of Chinese teachers and students. In China, especially in mainland China, both English teachers and learners attach great importance to grammar and stress precision, particularly at the sentence level. In addition, a cloze test is an inevitable part of any kind of English test in mainland China, CET-4 and CET-6 in particular, so preparing for such exams requires one to be acquainted with the relevant skills in doing such an exercise. The effect of formal schemata on reading comprehension might be overridden by the effect of those techniques. The differences in teaching systems between Hong Kong and mainland China, together with the difference in the subjects between this study and Sharp's [2002], correspondingly, led to different findings. Hong Kong was a colony of Britain; accordingly, its current teaching system (including the teaching aim, method and strategies) was strongly influenced by Britain. A wide gap exists between the teaching systems (including the policies on educational administration, the language of instruction, the allocation of funds, the examination system, the system of academic awards and the recognition of educational qualifications) of Hong Kong and mainland China [Bray 1997]. For example, English is the medium in class in



Sharp's study [2002] while Chinese is the instruction language in this study except in English class, which inevitably exerted a certain effect on the results of the two studies. Furthermore, the subjects in Sharp's study [2002] were Hong Kong Chinese school children (mean age 14.1) while the subjects in this study are sophomores in a mainland China university. They differ greatly in terms of age, language environment, etc.

However, in the light of recall protocol, both quantitatively and qualitatively, significant differences can be observed among different text types. This result can be explained well by the schemata theory. The last two types of formal schema — comparison and contrast, along with problem-solution — have an additional link of relationship over the descriptive schema. Accordingly, recall of information conveyed by the last two formal schemata, which offer extra linkage, should be better than that of the descriptive schema only if the subjects have a great ability of identifying the formal schemata. This is because Foo's study [1989] suggests that rhetorical structures (formal schemata) which help information recall are not necessarily easy to recognize. The subjects of this study were brought up in mainland China's education system which is exam-oriented, and they are strongly influenced by it. Besides, the majority of important English writing tests in mainland China such as CET-4, which is of vital importance to college students like the subjects in this study, is often exposition/argumentation writing and is seldom description writing. Therefore, both teachers and students take great effort to analyze and to practice the rhetorical patterns of such writings. This can explain why they recall better in comparison and contrast and in problem-solution than in description.

The difference between results from the cloze test and the recall test of the present study can be viewed from another perspective — being attributed to the inherent difference between cloze and recall tests. Some researchers [Alderson 1979; Lado 1986; Markham 1985] claim that the cloze test in general is more a test of linguistic skills (*e. g.*, grammatical and lexical knowledge) than of reading comprehension because cloze items are often based on cues from the immediate environment around the blank rather than on information from the whole text. In contrast, a recall test involving production of information may additionally require the selection and coordination of ideas and impressions, formulation, and ordering of remembered information [Meyer 1984]. The production skills that are required by recall tasks may affect both the quantity and the quality of information students recall [Johnston 1983]. According to Johnston's description of the cognitive requirements of recall tasks, the reader must understand and store the information, must be able to retrieve it on demand, and must decide on a starting point, a path through the information. The above discussion boils down to one point that, in comparison with cloze test, a recall test relies more on text's rhetorical structure, which might be the cause for different results from the cloze and recall tests being obtained.

## 5. Conclusion

In this thesis, the author has mainly studied the effect of formal schemata on reading comprehension by making a comparison of the reading scores of three schematically different texts by three groups of subjects, respectively. The three text types are description, comparison and contrast, and problem-solution. The results indicate that significant differences do exist among the text types both in terms of the quantity and quality of the recall protocol with the highly structured schema — problem-solution scoring the highest and the loosely controlled schema — description scoring the lowest. However, significant differences are not found among these three types in the cloze test.

This study has called attention to the formal schemata in written communication. Traditionally, teaching of writing emphasizes the instruction of new words and syntactic structures so students are unlikely to create communication problems at the sentence level, while producing many essays described by western scholars as written in a roundabout way, being flashy, and lacking consistence and logic [Coe and Hu 1989]. As this research demonstrates, texts written with clear structure and logical connection yield efficient communication. Therefore, it is the responsibility on the part of English writing teachers to get students acquainted with the practice of embodying the particular formal schema in their specific writing task.

There are several limitations to the present study due to time and other resources. The major limitation of the study was the sample size. In this study, the sample size is not large enough. With the small number of participants, the experiment was conducted in a small scale; the data thus collected may not be large enough for statistically significant generalization. Thus, conclusions are drawn within this context. Another limitation is the text type. The text type studied only includes description, comparison-contrast, and problem-solution. Future research should consider exploring other text types and using larger samples.

## References

- Aitken, K. G., "Problems in Cloze Testing Re-examined," *TESL Reporter*, 8, 1975, pp. 2.
- Alderson, J. C., "The Cloze Procedure and Proficiency in English as a Foreign Language," *TESOL Quarterly*, 13(2), 1979, pp. 219-223.
- Bachman, L. F., "Performance on Cloze Tests with Fixed-ratio and Rational Deletions," *TESOL Quarterly*, 19(3), 1985, pp. 535-556.
- Bormuth, J. R., "Cloze Test Readability: Criterion Reference Scores," *Journal of Educational Measurement*, 5(3), 1968, pp. 189-196.
- Bransford, J. D., and J. J. Franks, "The Abstraction of Linguistic Ideas," *Cognitive Psychology*, 2(4), 1971, pp. 331-350.

- Bray, M., "Education and Colonial Transition: The Hong Kong Experience in Comparative Perspective," *Comparative Education*, 33(2), 1997, pp. 157-169.
- Brown, J. D., "Relative Merits of Four Methods for Scoring Cloze Tests," *Modern Language Journal*, 64(3), 1980, pp. 311-317.
- Carrell, P. L., "Culture-specific Schemata in L2 Comprehension," *Selected Papers from the Ninth Illinois TESOL/BE Annual Convention, the First Midwest TESOL Conference*, ed. by R. Orem & J. Haskell, Illinois TESOL/BE, Chicago, 1981, pp. 123-132.
- Carrell, P. L., "Cohesion Is Not Coherence," *TESOL Quarterly*, 16(4), 1982, pp. 479-488.
- Carrell, P. L., "Some Issues in Studying the Role of Schemata or Background Knowledge in Second Language Comprehension," *Reading in a Foreign Language*, 1(1), 1983, pp. 81-92.
- Carrell, P. L., "The Effects of Rhetorical Organization on ESL Readers," *TESOL Quarterly*, 18(3), 1984, pp. 441-469.
- Carrell, P. L., "Introduction: Interactive Approaches to Second Language Reading," *Interactive Approaches to Second Language Reading*, ed. by P. L. Carrell, J. Devine, & D. E. Eskey, Cambridge University Press, New York, 2006, pp. 1-7.
- Carrell, P. L., and J. C. Eisterhold, "Schema Theory and ESL Reading Pedagogy," *TESOL Quarterly*, 17(4), 1983, pp. 553-573.
- Coe, R. M., and S. Hu, "A Preliminary Study of Contrastive Rhetorical Patterns in English and Chinese," *Journal of Foreign Languages*, (2), 1989, pp. 40-46.
- van Dijk, T., and W. Kintsch, *Strategies of Discourse Comprehension*, Academic Press, New York, 1983.
- Farhady, H., and M. N. Keramati, "A Text Driven Method for the Deletion Procedure in Cloze Passages," *Language Testing*, 13(2), 1996, pp. 191-207.
- Foo, R. W. K., "A Reading Experiment with L2 Readers of English in Hong Kong: Effects of the Rhetorical Structure of Expository Texts on Reading Comprehension," *Hong Kong Papers in Linguistics and Language Teaching*, 12, 1989, pp. 49-62.
- Goh, S. T., "The Effects of Rhetorical Organization on Expository Prose on ESL Readers in Singapore," *RELC Journal*, 21(2), 1990, pp. 1-11.
- Hudson, T., "The Effects of Induced Schemata on the 'Short Circuit' in L2 Reading: Non-decoding Factors in L2 Reading Performance," *Language Learning*, 32(1), 1982, pp. 1-31.
- Johnson, R. E., "Recall of Prose as a Function of the Structural Importance of the Linguistic Units," *Journal of Verbal Learning and Verbal Behaviour*, 9(1), 1970, pp. 12-20.
- Johnson, P., "Effects on Reading Comprehension of Language Complexity and Cultural Background of a Text," *TESOL Quarterly*, 15(2), 1981, pp. 169-181.
- Johnson, P., "Effects on Reading Comprehension of Building Background Knowledge," *TESOL Quarterly*, 16(4), 1982, pp. 503-516.

- Johnston, P. H., *Reading Comprehension Assessment: A Cognitive Basis*, International Reading Association, Newark, 1983.
- Jonz, J., "Another Turn in the Conversation: What Does Cloze Measure," *TESOL Quarterly*, 24(1), 1990, pp. 61-83.
- Kaplan, R. B., "Cultural Thought Patterns in Intercultural Education," *Language Learning*, 16(1), 1966, pp. 1-20.
- Lado, R., "Analysis of Native Speaker Performance on a Cloze Test," *Language Learning*, 3(2), 1986, pp.130-146.
- Markham, P. L., "The Rational Deletion Cloze and Global Comprehension in German," *Language Learning*, 35(3), 1985, pp. 423-430.
- Meyer, B. J. F., "Organizational Aspects of Text: Effects on Reading Comprehension and Applications for the Classroom," *Promoting Reading Comprehension*, ed. by J. Flood, International Reading Association, Newark, 1984, pp. 113-138.
- Meyer, B. J. F., and R. Freedle., *The Effects of Different Discourse Types on Recall*, Educational Testing Service, Princeton, 1979.
- Meyer, B. J. F., D. Brandt and G. Bluth., "Use of Top-Level Structure in Text: Key for Reading Comprehension of Ninth-grade Students," *Reading Research Quarterly*, 15(1), 1980, pp. 72-102.
- Morrow, L. M., "Retelling Stories as a Diagnostic Tool," *Re-examining Reading Diagnosis*, ed. by S. M. Glazer, L. W. Searfoss, & L. M. Gentile, International Reading Association, Newark, 1988, pp. 128-149.
- Oller, J. W., "Cloze Tests of Second Language Proficiency and What They Measure," *Language Learning*, 23(1), 1973, pp. 105-118.
- Peregoy, S. F., and O. F. Boyle, "English Learners Reading English: What We Know, What We Need to Know," *Theory into Practice*, 39(4), 2000, pp. 237-247.
- Plaister, T., "Reading Instruction for College Level Foreign Students," *TESOL Quarterly*, 2(3), 1968, pp. 164-168.
- Porter, D., "Cloze Procedure and Equivalence," *Language Learning*, 28(2), 1978, pp. 333-340.
- Rivers, W., *The Psychologist and the Foreign-language Teacher*, University of Chicago Press, Chicago, 1964.
- Rivers, W., *Teaching Foreign Language Skills*, University of Chicago Press, Chicago, 1968.
- Roller, C. M., "Commentary: The Interaction of Knowledge and Structure Variables in the Processing of Expository Prose," *Reading Research Quarterly*, 25(2), 1990, pp. 79-89.
- Rumelhart, D. E., "Toward an Interactive Model of Reading," *Attention and Performance*, Vol. 6, ed. by S. Dornic, Academic Press, New York, 1977a, pp. 573-603.
- Rumelhart, D. E., "Understanding and Summarizing Brief Stories," *Basic Processes in Reading: Perception and Comprehension*, ed. by D. LaBerge & S. Samuel, Erlbaum, Hillsdale, 1977b, pp. 265-303.

- Rumelhart, D. E., "Schemata: The Building Blocks of Cognition," *Theoretical Issues in Reading Comprehension*, ed. by R. J. Spiro, B. C. Bruce, & W. F. Brewer, Erlbaum, Hillsdale, 1980, pp. 33-58.
- Rumelhart, D. E., and A. Ortony, "The Representation of Knowledge in Memory," *Schooling and the Acquisition of Knowledge*, ed. by R. C. Anderson, R. J. Spiro, & W. E. Montague, Erlbaum, Hillsdale, 1977, pp. 99-135.
- Salvia, J., and C. Hughes, *Curriculum-based Assessment: Testing What Is Taught*, Macmillan Publishing Company, New York, 1990.
- Sanford, A. J., and S. C. Garrod, *Understanding Written Language*, Wiley, New York, 1981.
- Sharp, A., "Chinese L1 Schoolchildren Reading in English: The Effects of Rhetorical Patterns," *Reading in a Foreign Language*, 14 (2), 2002, pp. 1-20.
- Steffensen, M. S., C. Joag-dev, and R.C. Anderson, "A Cross-cultural Perspective on Reading Comprehension," *Reading Research Quarterly*, 15 (1), 1979, pp. 10-29.
- Stubbs, J. B., and G. R. Tucker, "The Cloze Test as a Measure of English Proficiency," *Modern Language Journal*, 58(5/6), 1974, pp. 239-241.
- Talbot, D., P. Ng, and A. Allan, "Hong Kong Students Reading Expository Prose: Replication of the Effects of Rhetorical Organization on ESL Readers by Patricia Carrell," *Working Papers of the Department of English, City Polytechnic of Hong Kong*, 3(1), 1991, pp. 52-63.
- Xiao, Liming, *English-Chinese Comparative Studies & Translation*, Shanghai Foreign Language Education Press, Shanghai, 2001.
- Yorio, C. A., "Some Sources of Reading Problems for Foreign Language Learners," *Language Learning*, 21(1), 1971, pp. 107-115.

## **Appendix A: The texts used in the experiment**

### **Passage 1 (Description)**

With the rapid expansion of cities, a huge number of people flocked there, and therefore, the demand of houses exceeds the supply of houses. Quite a lot people have no house to live in. Modern science and technology have made it possible for people to build high buildings and to supply water, electricity and elevators for people living in them. It is also proposed that government should open up underground housing area. High buildings can be set up on the vast pieces of land, on which old buildings of one, two or three stories have been pulled down. In this way a great number of people can get released from this trouble. Opening up an underground housing area is unrealistic because it needs much more money and takes a much longer time to carry out. Besides, since workers have to work under streets and other high buildings in order to build underground houses, it is rather dangerous. In the meantime, thousands of people are waiting anxiously, so enhancing the efficiency of their work not only concerns the happy life of many families but also bears upon the establishment of a harmonious society. (192 words)

### **Passage 2 (Comparison and Contrast)**

With the rapid expansion of cities, a huge number of people flocked there, which could not offer enough housing. Building high buildings and opening up underground housing areas are the two suggestions proposed to alleviate housing shortage. **HOWEVER, THE FORMER IS MORE PRACTICAL THAN THE LATTER.**

**BUILDING HIGH BUILDINGS IS EASIER TO CARRY OUT AND LESS EXPENSIVE.** Modern science and technology have made it possible for people to build high buildings and to supply water, electricity and elevators for people living in them. Furthermore, high buildings can be set up on the vast pieces of land, on which old buildings of one, two or three stories have been pulled down. In this way a great number of people's housing problems can be solved in a relatively short time. **IN CONTRAST,** opening up an underground housing area is unrealistic. It needs much more money and takes a much longer time to carry out. Besides, since workers have to work under streets and other high buildings in order to build underground houses, it is very dangerous.

Though both methods can fulfill the task, **BUILDING HIGH BUILDINGS IS MORE PRACTICAL THAN OPENING UP UNDERGROUND HOUSING AREAS.** (193 words)

### **Passage 3 (Problem-solution)**

With the rapid expansion of cities, a huge number of people flocked there, and **THE HOUSING PROBLEM** in big cities became one of the most serious of all the great problems which face us at the present time. And it is very urgent for us to take effective steps **TO**

SOLVE THIS PROBLEM.

THERE IS MORE THAN ONE WAY THAT PEOPLE SUGGEST TO SOLVE THIS PROBLEM. We could build high buildings. Modern science and technology have made it possible for people to build high buildings and to supply water, electricity and elevators for people living in them. Furthermore, high buildings can be set up on the vast pieces of land, on which old buildings of one, two or three stories have been pulled down. IN THIS WAY A GREAT NUMBER OF PEOPLE'S HOUSING PROBLEMS CAN BE SOLVED IN A RELATIVELY SHORT TIME. ANOTHER SOLUTION TO THIS PROBLEM is to open up an underground housing area. This method demands more money and a much longer time. Besides, workers have to work under streets and other high buildings in order to build underground houses.

If the two methods are adopted, THE HOUSING PROBLEM IS SURE TO BE SOLVED.  
(196 words)

**Note:** The capitalized words explicitly signal the discourse type of each passage. The interrelationship of the components is unstructured in the description text, with no clear relationships being evident and therefore, there is no clear textual signals in that text.

## Appendix B: Sample Cloze (based on 6th word deletion)

### Problem-solution

*With the rapid expansion of cities, a huge number of people flocked there, and the housing problem in 1. \_\_\_\_\_ (big) cities became one of the 2. \_\_\_\_\_ (most) serious of all the great 3. \_\_\_\_\_ (problems) which face us at the 4. \_\_\_\_\_ (present) time. And it is very 5. \_\_\_\_\_ (urgent) for us to take effective 6. \_\_\_\_\_ (steps) to solve this problem.*

There 7. \_\_\_\_\_ (is) more than one way that 8. \_\_\_\_\_ (people) suggest to solve this problem. 9. \_\_\_\_\_ (We) could build high buildings. Modern 10. \_\_\_\_\_ (science) and technology have made it 11. \_\_\_\_\_ (possible) for people to build high 12. \_\_\_\_\_ (buildings) and to supply water, electricity 13. \_\_\_\_\_ (and) elevators for people living in 14. \_\_\_\_\_ (them). Furthermore, high buildings can be 15. \_\_\_\_\_ (set) up on the vast pieces 16. \_\_\_\_\_ (of) land, on which old buildings 17. \_\_\_\_\_ (of) one, two or three stories 18. \_\_\_\_\_ (have) been pulled down. In this 19. \_\_\_\_\_ (way) a great number of people's 20. \_\_\_\_\_ (housing) problems can be solved in 21. \_\_\_\_\_ (a) relatively short time. Another solution 22. \_\_\_\_\_ (to) this problem is to open 23. \_\_\_\_\_ (up) an underground housing area. This 24. \_\_\_\_\_ (method) demands more money and a 25. \_\_\_\_\_ (much) longer time. Besides, workers have 26. \_\_\_\_\_ (to) work under streets and other 27. \_\_\_\_\_ (high) buildings in order to build 28. \_\_\_\_\_ (underground) houses.

If the two methods 29. \_\_\_\_\_ (are) adopted, the housing problem is 30. \_\_\_\_\_ (sure) to be solved.



## Appendix C

**Table 1. Idea Units and Their Hierarchical Arrangement of the Problem-solution Text**

Level of importance	Idea unit
1	With the rapid expansion of cities, a huge number of people flocked there,
3	and the housing problem in big cities became one of the most serious of all the great problems which face us at the present time.
3	And it is very urgent for us to take effective steps to solve this problem.
3	There is more than one way that people suggest to solve this problem.
2	We could build high buildings.
1	Modern science and technology have made it possible for people to build high buildings and to supply water,
1	electricity
1	and elevators for people living in them.
1	Furthermore, high buildings can be set up on the vast pieces of land,
1	on which old buildings of one, two or three stories have been pulled down.
2	In this way a great number of people's housing problems can be solved in a relatively short time.
2	Another solution to this problem is to open up an underground housing area.
1	This method demands more money
1	and a much longer time.
1	Besides, workers have to work under streets
1	and other high buildings in order to build underground houses.
3	If the two methods are adopted,
3	the housing problem is sure to be solved.

Note. Level of importance: 3=main generalization; 2=supporting generalization; 1=supporting detail.

**Appendix D**

**Questionnaire**

Name: \_\_\_\_\_ Age: \_\_\_\_\_ Sex: \_\_\_\_\_

Hometown: \_\_\_\_\_

Major: \_\_\_\_\_

Employment Objective: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Employment Experiences: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# A Cross-Linguistic Study of Voice Onset Time in Stop Consonant Productions

Kuan-Yi Chao\* and Li-mei Chen\*

## Abstract

This study examines voice onset time (VOT) for phonetically voiceless word-initial stops in Mandarin Chinese and in English, as spoken by 11 Mandarin speakers and 4 British English speakers. The purpose of this paper is to compare Mandarin and English VOT patterns and to categorize their stop realizations along the VOT continuum. As expected, the findings reveal that voiceless aspirated stops in Mandarin and in English occur at different places along the VOT continuum and the differences reach significance. The results also suggest that the three universal VOT categories (*i.e.* long lead, short lag, and long lag) are not fine enough to distinguish the voiceless stops of these two languages.

**Keywords:** Voice Onset Time (VOT), Voiceless Stops

## 1. Introduction

Over the past few decades, beginning with Lisker and Abramson's [Lisker and Abramson 1964] study, a considerable number of studies have investigated voicing contrasts in stops by the use of voice onset time (VOT). VOT has come to be regarded as one of the most important methods for examining the timing of voicing in stops (especially in word-initial position) and has been applied in studies of many languages. However, only a few attempts have been made to examine VOT patterns in Mandarin so far. Three universal categories of phonetically voiceless stops are generally recognized [Lisker and Abramson 1964], and Mandarin and English occupy the same place along the VOT continuum according to this general categorization. Nevertheless, differences between voiceless stops in Mandarin and English do exist. Since no existing studies compare the VOT patterns of these two languages, the aim of the present study is to provide a comparison of phonetically voiceless stops in Mandarin and in English, and to pinpoint the differences between their VOT patterns.

---

\*Department of Foreign Languages and Literature, National Cheng Kung University, 1 University Rd., Tainan, Taiwan. Telephone: (06)2757575 ext. 52231

E-mail: leemay@mail.ncku.edu.tw

The author for correspondence is Li-mei Chen.

## 2. Literature Review

Voicing contrast in stops has been widely discussed in phonetics and phonology. Voice Onset Time (VOT), the acoustic cue used to measure the timing of voicing, was first described by Lisker and Abramson in their well-known cross-language study of voicing in initial stops in 11 languages. According to Lisker and Abramson, voice onset time serves as a device for ‘separating the stop categories of a number of languages in which both the number and phonetic characteristics of such categories are said to differ.’ The authors also indicate that the measure of VOT is found to be highly effective in separating phonemic categories, such as voiced and voiceless, although the languages under study differ both in the number of those categories and their phonetic features. For example, in English, the minimal pair ‘pan’ and ‘ban’ can only be distinguished by voicing contrast. Thus, it can be seen that VOT plays an important role in differentiating voiced from voiceless stops, especially for lexical purposes.

### 2.1 VOT Definition

Lisker and Abramson’s study defined VOT as ‘the time interval between the burst that marks release of the stop closure and the onset of quasi-periodicity that reflects laryngeal vibration’ ([Lisker and Abramson 1964: 422]), and used the concept to examine word-initial stops in 11 languages. Since then, a considerable number of studies of many languages has been undertaken, including a report on VOT in 51 languages [Keating *et al.* 1983], and another more recent study on VOT in 18 languages [Cho and Ladefoged 1999].

Although VOT is now in widespread use for measuring the timing of voicing in stops, its role as a reliable measure to distinguish between voiced and voiceless stops has been brought into question. Bohn and Flege’s [Bohn and Flege 1993] findings suggest that VOT, the acoustic parameter of voicing contrasts in word-initial stops, may not be as important to the perception of stop voicing as is commonly supposed. Docherty [Docherty 1992] argues that voice onset time focuses narrowly on the timing of voicing in word-initial stops and does not take into account stops in word-final and word-medial positions. Caramazza, Yeni-Komshian, Zurif, and Carbone [Caramazza *et al.* 1973] also conclude that VOT is an ‘insufficient’ cue to the voicing contrast for French-English bilinguals.

On the other hand, some researchers argue that other acoustic cues play a role. For example, Klatt [Klatt 1975: 695] suggests that there are five equally important acoustic cues in English other than voice onset time, namely, low frequency energy in following vowels, burst loudness, fundamental frequency, pre-voicing, and segmental duration. Nevertheless, despite some research showing the limitations of voice onset time, it is still regarded as one of the most important acoustic parameters and has been used extensively in measuring word-initial stops.

## 2.2 VOT Category

Lisker and Abramson [Lisker and Abramson 1964: 388] examine 11 languages and classify them into three groups according to the number of stop categories each language contains. They also suggest that each stop category falls into one of three ranges,  $-125$  to  $-75$  ms,  $0$  to  $+25$  ms, and  $+60$  to  $+100$  ms, respectively (p. 403). Following Lisker and Abramson's categorization, both Mandarin and English fall into the two-category group of languages and occupy the same range along the VOT continuum, that is,  $0$  to  $+25$  ms for  $[p, t, k]$  and  $+60$  to  $+100$  ms for  $[p', t', k']$ .

However, this classification is too general to note the subtle variations between the two languages. Cho and Ladefoged [Cho and Ladefoged 1999: 223] go into further detail and classify the range for voiceless aspirated and unaspirated occlusives, concentrating particularly on velar stops across 18 languages. They distinguish four categories, which they name *unaspirated* (velar stops with a mean VOT of around 30 ms), *slightly aspirated* (with a mean VOT of around 50 ms), *aspirated* (with a mean VOT of around 90 ms), and *highly aspirated* (with a mean VOT of over 90 ms). According to Cho and Ladefoged [Cho and Ladefoged 1999: 223], Mandarin and English do not occupy the same place along the VOT continuum, especially for voiceless aspirated stops. Further discussion of VOT categories for Mandarin and English stops is in Section 2.4, below.

## 2.3 Effect on VOT

Voice onset time (VOT) is known to vary with place of articulation. The general principle is that, the further back the place of articulation, the higher the VOT values [Fischer-Jorgensen 1954; Lisker and Abramson 1964; Docherty 1992; Cho and Ladefoged 1999]. Cho and Ladefoged suggest several ways of explaining this principle, including explanations based on the laws of aerodynamics, articulator movement and differences in mass of articulators. According to the explanation based on the laws of aerodynamics, the main reason for velar stops having a longer VOT than alveolar or bilabial stops is the relative size of the supraglottal cavity behind the constriction. With the velar stop, greater air pressure builds up in the vocal tract because the supraglottal cavity becomes smaller and it takes longer for the pressure to fall at the beginning of the release phase.

In accordance with Lisker and Abramson's findings, velar stops have consistently higher VOT values than the other stops. Rochet and Fei's [Rochet and Fei 1991] study, which examines Mandarin stops, shows the same trend — velar stops consistently show the longest VOT. They also find that, in Mandarin voiceless aspirated stops (*i.e.*  $[p', t', k']$ ), the apical stop is correlated with slightly lower values than the labial one; this does not conform to the general agreement.

Vowel quality is another intrinsic effect which plays a crucial part in affecting voice

onset time. Although Lisker and Abramson claim that vocalic environment does not have a major influence on VOT, a number of reports have questioned their claim [Klatt 1975; Weismer 1979; Port and Rotunno, 1979]. Generally, it has been found that tense high vowels have longer VOTs than lax low vowels [Klatt 1975; Weismer 1979; Port and Rotunno 1979]. However, owing to language-specific variation, the correlation between voice onset time and vowel quality does not allow any definite conclusions.

Rochet and Fei state briefly that the mean VOT for both sets of Mandarin stops have greater values when they are followed by a high vowel /i/ or /u/ rather than the low vowel /a/. The study provides information on the phonetic features of VOICED and VOICELESS stops in Mandarin. (In this study, the upper-case forms — ‘VOICED’ and ‘VOICELESS’ — will be used to refer to stops’ phonological status, while the lower-case forms — ‘voiced’ and ‘voiceless’ — refer to their phonetic type.) More instrumental studies are needed in order to establish more complete and reliable Mandarin VOT patterns.

## 2.4 Mandarin and English Stops and VOT Patterns

As mentioned above, a sizable body of studies has been carried out to investigate the phonetic characteristics of voiced and voiceless stops in various languages using voice onset time as an important acoustic cue. Most existing studies concentrate on English [Lisker and Abramson 1964; Klatt 1975; Port and Rotunno 1979; Weismer 1979; Keating *et al.* 1983; Docherty 1992]. Other languages examined include Spanish [Lisker and Abramson 1964; Flege and Hammond 1982; Flege and Eefting 1987; Fellbaum 1996], French [Caramazza *et al.* 1973; Rochet *et al.* 1987], Arabic [Flege 1980; Khattab 2000] and Japanese [Shimizu 1990; Riney and Takagi 1999]. Among these investigations, there is very little data available on VOT for Mandarin word-initial stops. Thus, it does not appear that Mandarin VOT patterns have been examined extensively; to our knowledge, only Rochet and Fei have examined Mandarin Chinese.

This study will only discuss syllable initial stops, owing to the absence of stops in any other position in Mandarin Chinese. It is known that, in word-initial position, English VOICED stops are voiced or voiceless and unaspirated, and that VOICELESS stops are voiceless and aspirated [Keating *et al.* 1983; Keating 1984; Docherty 1992]. Although there are two possible phonetic implementations of English VOICED stops, Keating [Keating 1984: 43] indicates that ‘English divides up the VOT continuum with some lead values but mainly short lag vs. long lag.’ In Lisker and Abramson, VOT measurements occurring before the release burst are assigned negative values and called *voicing lead*, while VOT measurements occurring after the release burst are assigned positive values and called *voicing lag*. Lisker and Abramson [Lisker and Abramson 1964: 395] also provide two sets of values for English voiced stops (VOTs with lead and with short lag) and suggest that only a single type is

produced by each native speaker. Based on the distinction of Keating and Lisker and Abramson, English is described as having, in general, short lag and long lag VOT patterns.

In comparison with English, Mandarin shows less variation in implementation. All Mandarin stops are phonetically voiceless and are only differentiated by aspiration. According to data provided by Rochet and Fei, VOT duration for Mandarin [p', t', k'] ranges between 90 and 110 ms, while that of Mandarin [p, t, k] ranges between 10 and 25 ms (depending on the place of articulation). In Keating's study, Mandarin and English stops are classified as phonetically the same, and both fall into short lag vs. long lag patterns; however, there remain some subtle differences between the two languages.

Table 1 shows detailed measurements of English VOT means and ranges, including American English [adopted from Lisker and Abramson 1964] and British English [adopted from Docherty 1992]. The VOT pattern in Mandarin Chinese [adopted from Rochet and Fei 1991] is presented in Table 2. According to Cho and Ladefoged's definition of voiceless unaspirated stops, English [p, t, k] fall into the 'unaspirated' range with a VOT of under 30 ms. However, for voiceless aspirated stops (*i.e.* [p', t', k']) Mandarin occupies a 'highly aspirated' position along the VOT continuum, while English lies in the 'aspirated' region. The table also indicates that although [p', t', k'] in Mandarin and English fall into two categories, they are not completely different because the VOT ranges in these two languages overlap.

**Table 1. VOT means and ranges in English.**

	Lisker and Abramson 1964 (AE)		Docherty 1992 (BE)	
	Mean	Range	Mean	Range
<b>p'</b>	58	20 -120	42	10 - 80
<b>t'</b>	70	30 -105	64	30 - 110
<b>k'</b>	80	50 -135	62	30 - 150
<b>p</b>	1	0 - 5	15	0 - 50
<b>t</b>	5	0 - 25	21	0 - 50
<b>k</b>	21	0 - 35	27	10 - 60

(AE=American English; BE=British English; all measurements are in milliseconds (ms). Note: In this study, based on Keating [Keating 1984] and Lisker and Abramson's [Lisker and Abramson 1964] distinction, [p', t', k'] is used for referring to voiceless aspirated stops while [p, t, k] is used as voiceless unaspirated stops in phonetic distinctions.)

**Table 2. VOT means for voiceless aspirated stops in Mandarin.**

Rochet and Fei 1991	
	Mean ( in ms)
<b>p'</b>	99.6
<b>t'</b>	98.7
<b>k'</b>	110.3

(Note: Rochet and Fei provide the mean VOT for voiceless aspirated [p', t', k'] but not for voiceless unaspirated [p, t, k], nor do they provide VOT ranges.)

### 3. Methodology

#### 3.1 Aims of the Experiment

As mentioned above, a few instrumental studies have examined VOT in Mandarin. No existing studies even attempt to compare VOT patterns in Mandarin and English. Therefore, the present experiment examines subtle differences in VOT production between the two languages and seeks to determine whether Mandarin and English stop realizations occupy the same place along the VOT continuum.

#### 3.2 Linguistic Material

Stops in Mandarin Chinese occur only in the initial position. The present experiment examines two sets of phonetically voiceless stops, that is, aspirated [p', t', k'] and unaspirated [p, t, k]. Each of the stops is followed by three peripheral vowels in turn: two high vowels, /i/ and /u/, and one low vowel, /a/, thus giving a total of three variations for each stop. Two exceptions to this are the velar stops [k'] and [k], as no Mandarin lexical items exist for the sequences [k'] or [k] followed by the high front vowel /i/. All the words used are real words.

Unlike English, every character in Mandarin is correlated with monosyllabic sounds only, all of which can exist independently. However, two or more characters usually stand side by side to form a 'word' which is more complete and more meaningful. Take 'da ying' for example, the sound 'da' can have various meanings; however, 'da ying' means 'to promise' and is easy for experiment participants to understand. Thus, in the present experiment, compound words are used rather than single characters (as in previous experiments, *e.g.*, Rochet and Fei 1991) because they make more sense to the subjects, and because compound words fit the testing format better than single characters. The inventory of all stimuli words is listed in Appendix A.

The following procedure was followed to create English and Mandarin word lists. First, the word-lists were oriented to the Mandarin Chinese lexicon in order to obtain comparable material in both English and Mandarin. Thus, all the English stops examined here are in the word-initial position and the target words contain two high vowels (/i/ and /u/) and one low vowel (/a/). Velar stops [k'] and [k] followed by the high front vowel /i/ are not included as they do not occur in Mandarin. Secondly, disyllabic words were chosen rather than monosyllabic materials because the Mandarin word-list consists of pairs of characters put together to make target words. Every effort was made to find Mandarin and English words in which the stops were followed by 'similar' vocalic and consonantal contexts; however, it was very difficult to create a list of homophones across the two languages owing to phonetic and phonological differences.



### **3.3 Participants**

11 Mandarin speakers participated in this experiment: all were females aged between 19 and 35 years (mean=27). All of the participants were born in Taiwan, where Mandarin Chinese is the predominant language. None of them had a marked regional accent, although they were raised in different areas of Taiwan. All the Mandarin participants were college students. Four female native speakers of British English participated in this experiment as a control group. The native English speakers were older (mean=32) than the native speakers of Mandarin. All were born and raised in the UK. The Mandarin speakers participated in the Mandarin-based test, while the English speakers participated in the English-based test.

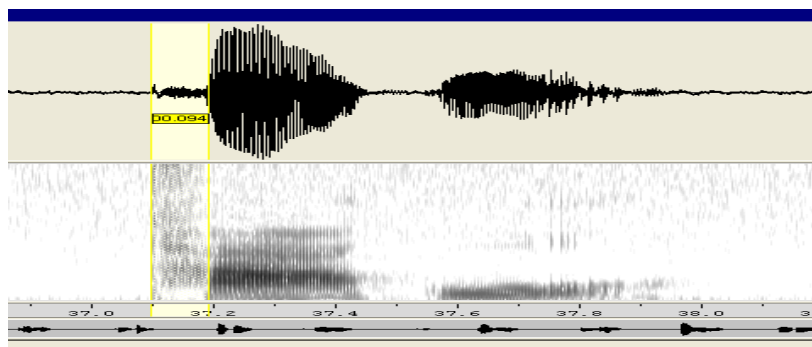
### **3.4 Procedures**

In order to ensure the target sounds were not predictable for any participants who might have linguistics backgrounds, all the speech materials, both Mandarin and English, were randomized. Target words were equally divided into two subgroups in each language due to recording limitations. Two preparation words were added to each group, as the first and last items, in order to allow participants to practice before the target words being recorded. One by one, participants were recorded in a sound booth, and their readings recorded straight into a computer. The participants were asked to start reading each word when they were ready; the list of words was read five times in a row at a comfortable speed. The first iteration was not used for analysis. This was to ensure that participants were familiar with the entire recording procedure before they produced a recording which would actually be analyzed. All speakers were given instructions in English, and Mandarin participants also received instructions in Mandarin. The participants were told that the object of the experiment was to examine speech, but they were not told that their production of specific sounds (*i.e.* stops) would be assessed.

### **3.5 Measurements and Analyses**

Acoustic measurements of the speech material were made using Wavesurfer software. Following Lisker and Abramson [Lisker and Abramson 1964: 422], VOT was measured as the interval between the beginning of the release burst and the onset of quasi-periodicity that reflects glottal vibration in F1 in the following vowel. Spectrographic measurements were taken of the VOTs of Mandarin initial stops in a total number of 800 tokens. Mean VOT values were calculated for the stops produced by each participant. A few measurements were missing owing to discarded tokens or imperceptible release bursts which made it difficult to measure VOT. The analysis involved displaying two panels (spectrogram and waveform) on separate portions of the screen, and using a manually controlled cursor for durational measurements, as shown in Figure 1. Measurement reliability was assessed by re-measuring three randomly selected stops in each group from duplicate spectrograms. The average

difference between the two measurements was 2.2 ms, with a range of 0-5 ms. T-tests were used for comparison of results and calculation of statistical significance. In this study, the p values which were less than 0.05 were reported to be significant.



*Figure 1. Example of a test word “ta guo” (踏過) shown in the acoustic waveform (top panel) and spectrogram (bottom panel).*

## 4. Results

### 4.1 Overall Results for Mandarin VOT

#### 4.1.1 VOT Means and Distribution

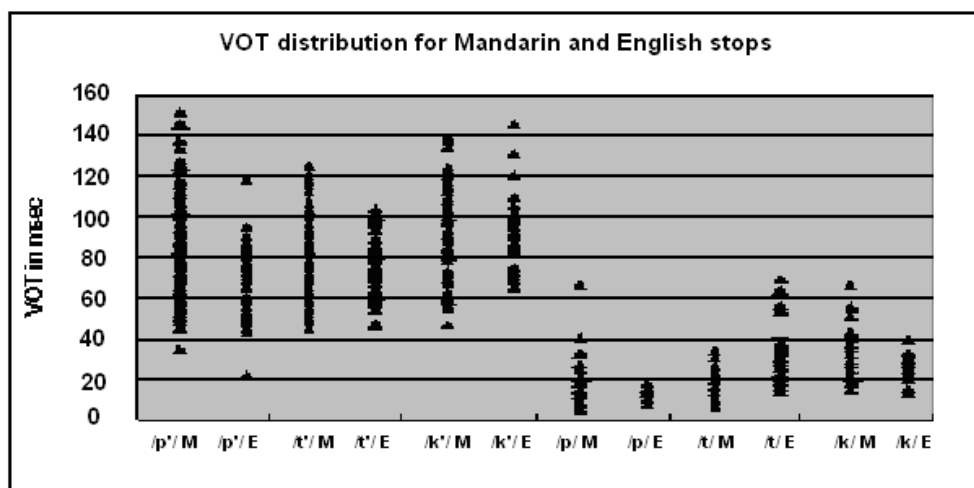
Table 3 presents the mean VOT values and ranges for each of the six Mandarin stops. It was found that the velar stops [kʰ] and [k] have higher VOT values than the other stops. This supports the findings in Lisker and Abramson [Lisker and Abramson 1964] in their cross-language study. The mean VOT value for [tʰ] is slightly lower than that for [pʰ], which does not conform to the general agreement that the further back the place of articulation, the longer the VOT. However, the difference between [pʰ] and [tʰ] does not reach significance. This concurs with Rochet and Fei’s finding [Rochet and Fei 1991] that the mean VOTs for [pʰ] and [tʰ] do not differ significantly from each other. T-tests reveal that the VOT value for the velar stop [kʰ] is significantly higher than those for [pʰ] and [tʰ].

Regarding the voiceless unaspirated stops, it is noticeable that the mean VOT for [k] is higher than that for [t], which in turn is higher than the VOT for [p]. Compared with voiceless aspirated stops, unaspirated [p, t, k] are in closer accord with the consensus reported in the literature. T-tests were carried out to determine if the differences between [p], [t], and [k] were statistically significant. The results show that the VOT for the velar stop is significantly higher than those of the other two stops, and there is a significant difference between [p] and [t] as well. Overall, the present study’s findings on Mandarin VOT patterns are in accordance with the study by Rochet and Fei, in which the authors found that the mean VOT for [tʰ] was slightly but not significantly lower than that for [pʰ].

**Table 3. VOT ranges and general means (in ms) for Mandarin stops.**

	<b>p'</b>	<b>t'</b>	<b>k'</b>	<b>p</b>	<b>t</b>	<b>k</b>
<b>Min</b>	35	45	50	7	7	15
<b>Max</b>	147	123	138	65	33	65
<b>General means</b>	82	81	92	14	16	27

VOT distribution for all stops in Mandarin and English is shown in Figure 2. As the figure indicates, the VOT ranges for voiceless aspirated [p', t', k'] are 35-147 ms, 45-123 ms and 50-138 ms, respectively. However, it is clear that the VOT distribution for [p'] tends to center in the range from 45 to 130 ms, and that for [k'] is concentrated in the range from 58 to 125 ms. As for voiceless unaspirated stops, the VOT ranges for [p, t, k] are 7-65 ms, 7-33 ms and 15-65 ms, respectively. VOT distribution is centered in the range of 7-28 ms for [p] and 15-50 ms for [k]. It is interesting to note that the ranges for [t'] and [t] are narrower than those for the other stops; that is, the bilabials and the velars allow more variation than [t'] and [t].



**Figure 2. VOT distribution for all stops in Mandarin (M) and English (E).**

### 4.1.2 Vowel Context

As mentioned above, vowel quality is another important factor influencing voice onset time. It is now widely accepted that tense high vowels are correlated with longer VOTs than lax low vowels [Klatt 1975; Weismer 1979; Port and Rotunno 1979]. Table 4 shows the mean VOTs for Mandarin stops followed by three peripheral vowels /i, a, u/. It should be noted that there are no lexical items with the sound sequences of /ki/ or /k'i/ in Mandarin. As shown in the table, all the stops have longer VOTs when they are followed by a high vowel, /i/ or /u/, than by the low vowel /a/. A series of t-tests were carried out to determine if the differences are statistically significant in the VOT values between the stops when followed by different vowels. The results indicate that all the stops, except [t'] which does not reach significance,

have significantly longer VOTs when the following vowel is /i/ or /u/ than when it is /a/. The present results are in keeping with the correlation between VOT and vowel quality reported in Rochet and Fei's findings. To sum up, it may safely be assumed that Mandarin stops generally fit in with the assumption that tense high vowels contribute to longer VOTs in preceding stops than lax low vowels.

**Table 4. Mean VOT values (in ms) for Mandarin stops with three following vowels /i/, /a/ and /u/**

	p'	t'	k'	p	t	k
i	90	82	X	13	18	X
u	87	82	98	18	17	33
a	70	81	86	12	14	22

#### 4.2 Overall Results for English VOT

Table 5 shows the mean VOT values produced by the British English speakers. English mean VOTs with and without subject A are juxtaposed for comparison. As can be clearly seen, the place of articulation has an effect on VOT for [p', t', k'] but not for [p, t, k]. Regarding the mean VOTs by native English speakers with subject A, it is found that the mean VOT for aspirated velar [k'] is significantly higher than that for [t'], and for [p']. T-tests also reveal that there is a significant difference between [t'] and [p']. As for unaspirated stops, it is apparent that the mean VOTs for both [t] and [k] are significantly longer than that for [p]. It should also be noted that the mean VOT value for [t] is slightly higher than that for [k], which does not support the general agreement that velar stops are usually produced with longer VOTs. With respect to the mean VOTs without subject A, it shows almost the same trend with the one with subject A but differs in two places. First, the mean VOT values for all English stops, especially for [p', t', k'], are lower when subject A is excluded. One explanation for this may be that subject A is the only English-Mandarin bilingual among the subjects, and this is also reflected in the distribution (see results below). Secondly, VOT values for [k] is slightly higher than that for [t], which conforms more closely to the general principle, although the difference between [t] and [k] does not reach significance.

**Table 5. Mean VOTs for English stops by British English speakers with and without subject A.**

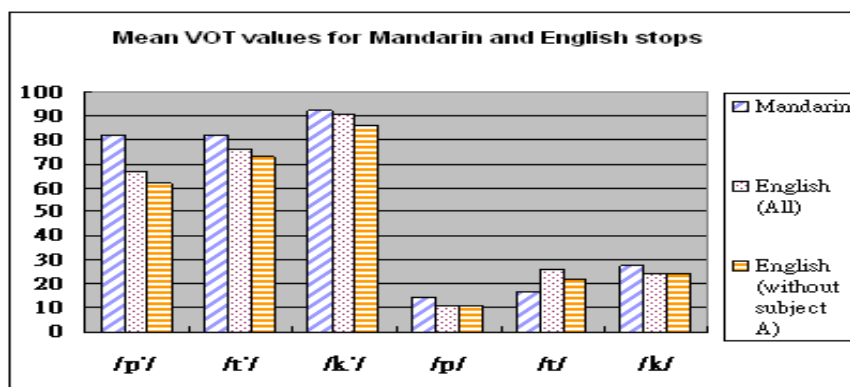
	/p'/	/t'/	/k'/	/p/	/t/	/k/
English (with subject A)	67	76	91	11	26	24
English (without subject A)	62	73	86	11	22	24

As for English VOT distribution, Figure 2 indicates that VOT ranges for [p', t', k'] are from 22 ms to 117 ms, 48 ms to 105 ms, and 65 ms to 145 ms, respectively; while the ranges

for [p, t, k] are 7-18 ms, 13-68 ms, and 13-40 ms, in that order. Some of the higher values, such as the highest value for [pʰ], the top two values for [kʰ], and the values from 52 ms to 68 ms for [t], were produced by subject A. Generally, VOT distribution in the present study fits the ranges reported in both Lisker and Abramson who examined American English, and Docherty [Docherty 1992], who concentrated on British English (their findings are provided in Table 1).

### 4.3 Comparing Mandarin and English VOT

VOT patterns in Mandarin and English are compared in terms of the mean VOT values and the distribution patterns. Figure 3 presents the mean VOTs for Mandarin and English stops. Chinese speakers generally produce longer VOTs than English speakers do, especially for voiceless aspirated stops. A series of t-tests were implemented to examine if the differences among /pʰ, tʰ, kʰ/ in both languages reach significance.

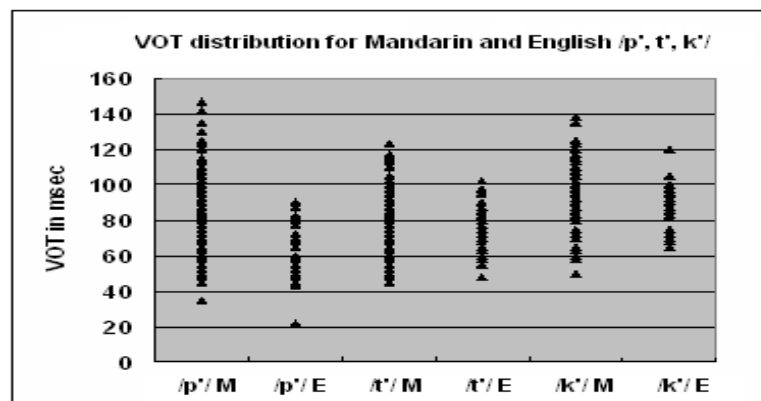


**Figure 3. Mean VOTs for Mandarin and English stops by Chinese speakers and English speakers with and without subject A**

Comparing the mean VOTs in Mandarin and English with subject A’s productions the results reveal that no comparisons reach significance although the figure indicates that VOT values for [pʰ, tʰ, kʰ] are longer in Mandarin than in English. However, when subject A is excluded from the group of English speakers, t-tests comparing Mandarin and English mean VOTs show significant results for all the aspirated stops although the differences between Mandarin and English VOT are subtle, not stark. It would thus be interesting to find out whether the L2 learners are aware of the subtle differences between the two languages and are capable of producing them authentically.

As Figure 3 shows, the mean VOT values for English [pʰ, tʰ, kʰ] are longer as the further back of the place of articulation. However, this does not apply to Mandarin [pʰ] and [tʰ]. The result accords with Rochet and Fei’s finding that the mean VOT durations for Mandarin [pʰ] and [tʰ] are close to each other, and [pʰ] always has a slightly higher value than [tʰ].

VOT distribution patterns for [p', t', k'] in Mandarin and in English are shown in Figure 4. Mandarin has higher VOT values than English; moreover, the ranges for Mandarin [p', t', k'] are wider than that for English ones. According to the definition for voiceless aspirated stops provided by Cho and Ladefoged [Cho and Ladefoged 1999], Mandarin occupies the 'highly aspirated' region along the VOT continuum while English falls into the 'aspirated' position. However, even if the two languages belong to the different categories, it does not mean that they occupy totally separate places along the VOT continuum. It can be clearly seen from the figure that the VOT ranges for Mandarin and English [p', t', k'] are considerably overlapped. The finding triggers a series of concerns about whether the native Chinese speakers will be aware of those subtle differences.



*Figure 4. VOT distribution patterns for Mandarin and English /p', t', k'/ produced by Chinese speakers (M) and English speakers (E) without subject A.*

## 5. Discussion

With respect to the comparison of VOT patterns in Mandarin and English, the results of this study corroborate the claim that significant differences exist between these two languages, especially for voiceless aspirated stops. It is known that all Mandarin and English stops are phonetically voiceless, and aspiration is becoming a common way of distinguishing homorganic pairs. According to Lisker and Abramson's categorization [Lisker and Abramson 1964], both Mandarin and English belong to the two-category group of languages, for which VOTs range from 0 to +25 ms for unaspirated [p, t, k] and from +60 to +100 ms for aspirated [p', t', k']. As expected, Mandarin and English follow short lag vs. long lag VOT patterns. However, the results of the study by Rochet and Fei [Rochet and Fei 1991] show that this classification is too general to observe the subtle differences between Mandarin and English. It is uninformative to compare unaspirated stops in the two languages as they occupy the same region along the VOT continuum and have close mean VOTs. As for aspirated stops, it may be

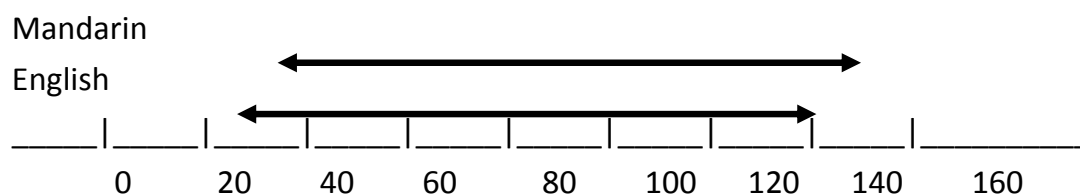
noted that Mandarin [p', t', k'] often have higher VOT values than English [p', t', k']. The findings showing higher mean VOTs in Mandarin are generally consistent with those reported by Rochet and Fei. However, the values obtained in this present study are slightly lower than Rochet and Fei because this experiment uses disyllabic words.

Aware that the three universal categories (*i.e.* long lead, short lag, and long lag) are not fine enough to distinguish the different regions along the long-lag VOT continuum, Cho and Ladefoged [Cho and Ladefoged 1999] suggest using four categories to define voiceless stops according to their degree of aspiration. In light of Cho and Ladefoged's categorization, it is clear that English [p', t', k'] fall into the 'aspirated' category with VOTs ranging from 50 ms to 90 ms. As for Mandarin, as shown in Table 6, all the VOT values of Mandarin [p', t', k'] in Rochet and Fei are over 90 ms, while in the present study the VOT values of the velar [k'] is over 90 ms. Therefore, Mandarin [p', t', k'] should be generally categorized as 'highly aspirated.' However, in this present study, Mandarin [p'] and [t'] might not fall into the 'highly' aspirated category. The results of the present study agree with Cho and Ladefoged's claim that more than three categories for aspirated stops need to be taken into consideration.

**Table 6. Mean VOT values (in ms) for Mandarin voiceless aspirated stops**

	Rochet and Fei 1991	The present study
	monosyllables	disyllables
p'	99.6	82.2
t'	98.7	81.9
k'	110.3	92.1

The study by Cho and Ladefoged used only VOT means for categorization. However, distribution is even more important. Figure 5 provides a simplified schematic representation of the places that Mandarin and English stops occupy along the VOT continuum. As shown in the figure, Mandarin [p', t', k'] occupies wider ranges and higher values than English. Moreover, it is difficult to draw a line between the two sets of VOT values for Mandarin and English [p', t', k'] due to the considerable overlap between the two distributions.



**Figure 5. Schematic representation of VOT ranges for English and Mandarin [p', t', k'] along the VOT continuum.**

Knowing that there is a great deal of overlap, it is interesting to find out whether the differences between Mandarin and English reach significance. The results of the present study show that all the differences for [p', t', k'] in both languages reach significance. It may thus imply that Mandarin and English occupy the same place along the VOT continuum for voiceless unaspirated stops, while the two languages belong to different categories for voiceless aspirated stops. Further studies are required using a greater number of monolingual subjects and examining [p', t', k'] in different contexts (in isolation and in sentences) in order to describe the VOT patterns of the Mandarin stop categories more accurately.

## 6. Conclusion

With the purpose of comparing voice onset time patterns in Mandarin and English as well as categorizing Mandarin stops by voice onset time, two general conclusions may be drawn from the present study. First, it is found that VOT patterns in the two languages are similar but not completely identical. Voiceless unaspirated [p, t, k], stop realizations in both languages occupy the same range: the short lag region along the VOT continuum. However, for voiceless aspirated [p', t', k'], Mandarin seems to fall into the 'highly aspirated' region along the VOT continuum, while English falls into the 'aspirated' region. The findings obtained in this study conform to the result reported by Rochet and Fei [Rochet and Fei 1991]. Moreover, the discrepancies between Mandarin and English are subtle due to the considerable overlap. Additional research with more monolingual subjects and a more natural setting (*i.e.* to obtain natural speech rather than elicited word lists) will be needed in order to accurately pinpoint the VOT patterns of Mandarin stops.

As for Mandarin stops categorization, the results suggest that the classification presented by Cho and Ladefoged [Cho and Ladefoged 1999] is more suitable than the three-way categorization [Lisker and Abramson 1964], especially for the language whose voicing contrast is the aspiration. Moreover, to discuss the distinction of voicing contrast, VOT is not the only parameter. Some researchers suggest other acoustic cues which share the equivalent importance with voice onset time to distinguish voicing contrast. These should be taken into account in future studies. Furthermore, two critical issues should be further examined: the irregular tendency for Mandarin [p', t', k'] and the underlying reasons for the differences between Mandarin and English VOT. The former can be discussed involving some possible factors, such as tones, vowel context, and place of articulation, while the latter can be detail explained by many aspects, for instance, different phonetic features or sound system in the two languages. Some researchers have claimed that tones play a role of being associated with VOT values [Liu *et al.* 2008]. The test stimuli used in the present experiment are not with the same tone; therefore, future studies should take it into consideration as well.



## References

- Bohn, O.S. and J.E. Flege, "Perceptual Switching in Spanish/English Bilinguals," *Journal of Phonetics*, 21(3), 1993, pp. 267-290.
- Caramazza, A., G. Yeni-Komshian, E. Zurif and E. Carbone, "The Acquisition of A New Phonological Contrast: The Case of Stop Consonants in French-English Bilinguals," *Journal of the Acoustical Society of America*, 54, 1973, pp. 421-428.
- Cho, T and P. Ladefoged, "Variation and Universals in VOT: Evidence from 18 Languages," *Journal of Phonetics*, 27, 1999, pp. 207-229.
- Docherty, G.J., *The Timing of Voicing in British English Obstruents*, Foris, New York, 1992.
- Fellbaum, M.L., "The Acquisition of Voiceless Stops in the Interlanguage of Second Language Learners of English and Spanish," *Proceedings of the 4<sup>th</sup> International Conference on Spoken Language Processing*, 3, 1996, Philadelphia, PA, pp. 1648-1651.
- Fischer-Jorgensen, E., "Acoustic Analysis of Stop Consonants," *Miscellanea Phonetica*, 1954, pp. 42-59.
- Flege, J.E., "Phonetic Approximation in Second Language Acquisition," *Language Learning*, 30(1), 1980, pp. 117-134.
- Flege, J.E. and W. Eefting, "Production and Perception of English Stops by Native Spanish Speakers," *Journal of Phonetics*, 15, 1987, pp. 67-83.
- Flege, J.E. and R.M. Hammond, "Mimicry of Non-Distinctive Phonetic Differences Between Language Varieties," *Studies in Second Language Acquisition*, 5(1), 1982, pp. 1-17.
- Keating, P.A., W. Linker, and M. Huffman, "Patterns in Allophone Distribution for Voiced and Voiceless Stops," *Journal of Phonetics*, 11, 1983, pp. 277-290.
- Keating, P.A., "Phonetic and Phonological Representation of Stop Consonant Voicing," *Language*, 60, 1984, pp. 286-319.
- Khattab, G., "VOT Production in English and Arabic Bilingual and Monolingual Children," *Leeds Working Papers in Linguistics and Phonetics*, 8, 2000, pp. 95-122.
- Klatt, D.H., "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," *Journal of Speech and Hearing Research*, 18, 1975, pp. 686-706.
- Lisker, L. and A.S. Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," *Word*, 20, 1964, pp. 384-422.
- Lisker, L. and A.S. Abramson, "Some Effects of Context on Voice Onset Time in English Stops," *Language Speech*, 10, 1967, pp. 1-28.
- Liu, H., M. Ng, M. Wan, S. Wang and Y. Zhang, "The Effect of Tonal Changes on Voice Onset Time in Mandarin Esophageal Speech," *Journal of Voice*, 22(2), 2008, pp. 210-218.
- Port, R.F. and R. Rotunno, "Relation Between Voice-Onset Time and Vowel Duration," *Journal of the Acoustical Society of America*, 66, 1979, pp. 654-662.
- Riney, T.J. and N. Takagi, "Global Foreign Accent and Voice Onset Time Among Japanese EFL Speakers," *Language Learning*, 49(2), 1999, pp. 275-302.

- Rochet, B.L. and Y. Fei, "Effect of Consonant and Vowel Context on Mandarin Chinese VOT: Production and Perception," *Canadian Acoustics*, 19(4), 1991, p. 105.
- Rochet, B., T. Nearey and M. Munro, "Effects of Voicing Place and Vowel Context on VOT for French and English Stops," *Journal of the Acoustical Society of America*, 81, 1987, S65.
- Shimizu, K., A Cross-Language Study of Voicing Contrasts of Stop Consonants in Asian Languages, PhD thesis, University of Edinburgh, 1990.
- Weismer, G., "Sensitivity of Voice-Onset-Time (VOT) Measures to Certain Segmental Features in Speech Production," *Journal of Phonetics*, 7, 1979, pp. 197-204.

**Appendix A**

**Table A.1. Inventory of Mandarin stops followed by two high vowels /i/ and /u/, and a low vowel /a/.**

	/p/	/t/	/k/	/p'/	/t'/	/k'/
/i/	逼迫 pi po	低聲 ti sheng	/	霹靂 p'i li	踢球 t'i qiu	/
/a/	八年 pa nian	答應 ta ying	嘎嘎 ka ga	趴下 p'a xia	踏過 t'a guo	卡鎖 k'a suo
/u/	布丁 pu ding	杜葳 tu wei	故宮 ku gong	撲倒 p'u dao	土匪 t'u fei	苦苓 k'u ling

**Table A.2. Inventory of English stops followed by two high vowels /i/ and /u/, and a low vowel /a/.**

	/p/	/t/	/k/	/p'/	/t'/	/k'/
/i/	beetle	decent	/	peeling	teacup	/
/a/	bath tub	darling	gagging	passion	tackle	castle
/u/	booting	duvet	google	poodle	tooth paste	cooling



# Data Driven Approaches to Phonetic Transcription with Integration of Automatic Speech Recognition and Grapheme-to-Phoneme for Spoken Buddhist Sutra

Min-Siong Liang\*, Ren-Yuan Lyu<sup>+</sup>, and Yuang-Chin Chiang<sup>#</sup>

## Abstract

We propose a new approach for performing phonetic transcription of text that utilizes automatic speech recognition (ASR) to help traditional grapheme-to-phoneme (G2P) techniques. This approach was applied to transcribe Chinese text into Taiwanese phonetic symbols. By augmenting the text with speech and using automatic speech recognition with a sausage searching net constructed from multiple pronunciations of text, we are able to reduce the error rate of phonetic transcription. Using a pronunciation lexicon with multiple pronunciations for each item, a transcription error rate of 12.74% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules derived manually from corrected transcription in a speech corpus. The PV rules can be categorized into two kinds: knowledge-based and data-driven rules. By incorporating the PV rules, an error rate of 10.56% could be achieved. Although this technique was developed for Taiwanese speech, it could easily be adapted to other Chinese spoken languages or dialects.

**Keywords:** Automatic Phonetic Transcription, Phone Recognition, Grapheme-to-Phoneme (G2P), Pronunciation Variation, Chinese Text, Taiwanese (Min-Nan), Dialect, Buddhist Sutra.

---

\* Dept. of Electrical Engineering, Chang Gung University, 259 Wen-Hwa 1<sup>st</sup> Rd., Kwei-Shan Tao-Yuan, Taiwan

E-mail: minsiong@gmail.com

<sup>+</sup> Dept. of Computer Science and Information Engineering, Chang Gung University, 259 Wen-Hwa 1<sup>st</sup> Rd., Kwei-Shan Tao-Yuan, Taiwan

E-mail: renyuan.lyu@gmail.com

<sup>#</sup> Institute of Statistics, National Tsing Hua University, Hsinchu, 101, Section 2 Kuang Fu Rd., 30013, Taiwan

E-mail: chiang@stat.nthu.edu.tw

## 1. Introduction

Automatic phonetic transcription is gaining popularity in the speech processing field, especially in speech recognition, text-to-speech, and speech database construction [Haeb-Umbach *et al.* 1995; Wu *et al.* 1999; Lamel *et al.* 2002; Evermann *et al.* 2004; Nanjo *et al.* 2004; Nouza *et al.* 2004; Sarada *et al.* 2004; Siohan *et al.* 2004; Soltau *et al.* 2005; Kim *et al.* 2005]. It is traditionally performed using two different approaches: an acoustic feature input method and a text input method. The former is the speech recognition task, or more specifically, the phoneme recognition task. The latter is the grapheme-to-phoneme (G2P) task. Both tasks, including phoneme recognition and G2P remain unsolved technology problems. The state-of-the-art speaker-independent (SI) phone recognition accuracy in a large vocabulary task is currently less than 80%, far from human expectations. Although the accuracy of G2P tasks seems much better, it relies on a “perfect” pronunciation lexicon and cannot effectively deal with pronunciation variation issues.

This problem becomes non-trivial when the target text is the Chinese text (漢字). The Chinese writing system is widely used in China and in East/South Asian areas including Taiwan, Singapore, and Hong Kong. Although the same Chinese character is used in different areas, the pronunciation may be very different. Therefore, they are mutually unintelligible and considered different languages rather than dialects by most linguists.

In this paper, we chose Buddhist Sutra (written collections of Buddhist teachings) as the target text processed in this research. Buddhism is a major religion in Taiwan (23% of the population) [IIP 2003]. The Buddhist Sutra, translated into Chinese text in a terse ancient style (古文), is commonly read in Taiwanese (Min-nan). Due to a lack of proper education, most people are not capable of correctly pronouncing all of the text. Besides, no qualified pronunciation lexicon exists and very few appropriately computational linguistic research projects have been conducted to support developing a G2P system.

Taiwanese uses Chinese characters as a part of the written form, with its own phonetic system differing greatly from Mandarin. This is in contrast to the case of Mandarin, where the problem of multiple pronunciations (MP) is less severe. A Chinese character in Taiwanese can commonly have a classic literate pronunciation (known as Wen-du-in, or “文讀音” in Chinese) and a colloquial pronunciation (known as Bai-du-in, or “白讀音” in Chinese) [Liang *et al.* 2004a]. In addition to MPs, Taiwanese also have pronunciation variation (PV) due to sub-dialectal accents, such as Tainan and Taipei accents. We use the term MPs to stress the fact that variation may cause more deterioration in phonetic transcription [Cremelie *et al.* 1999; Hain 2005; Raux 2004].

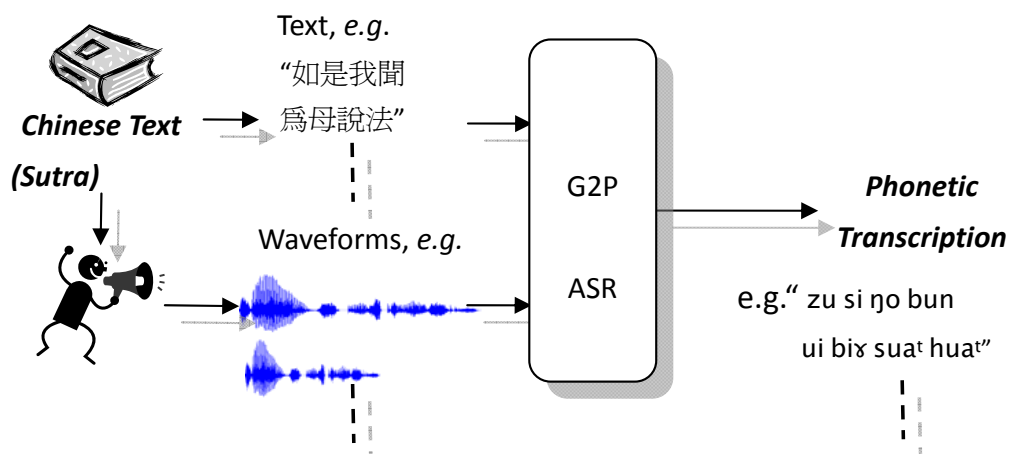
The traditional approach to transcribing Chinese Buddhist Sutra text is human dictation. A master monk or nun reads the text aloud, sentence by sentence. Then, some phonetic experts

transcribe the text manually. The manual transcription process is tedious and prone to errors. An example is given in Table 1 as follows [Chen 2006; Tripitaka *et al.* 2005].

**Table 1. An example of Transcription of Chinese Buddhist Sutra text into Taiwanese pronunciation, with English translation. The phonetic symbols used here are IPA followed by a digit representing one of several tone classes of the Taiwanese language.**

Chinese text of Buddhist Sutra	地藏菩薩本願經：如是我聞。 一時佛在忉利天，為母說法。
Transcription of Taiwanese Pronunciation	tè tsòŋ p'ò sá <sup>1</sup> pún gúan kíŋ: zù sī ŋó bunn í <sup>1</sup> sǐ hú <sup>1</sup> tsài t̃x̃ li tien, ui bĩx̃ súa <sup>1</sup> húa <sup>1</sup>
English translation in meaning	Sutra of Earth Treasure: Thus I heard, once the Buddha was in Dao Li Heaven to expound the Dharma to his mother

Since more transcribed Sutras are planned, we are interested in how G2P and ASR technology can help in this situation. Owing to the fact that human experts capable of phonetically transcribing the Sutra in Taiwanese are difficult to find, the first phonetically transcribed Sutra in Taiwanese did not appear until 2004 [Sik 2004a, 2004b]. As shown in Figure 1, our task is to discover which of them is actually pronounced when the Sutra text is segmented into a series of sentences and recorded by a senior master nun. Then, the output of transcription is formed in ForPA or Tongyong Pinin [Lyu *et al.* 2004]. These two phonetic symbol systems are well-designed in ASCII code and suitable for any learners with common understanding of the English phonetic system. This architecture is much easier for a person to use to record his/her reading of the text than acquiring a transcribing expert. For marginalized languages with serious MPs and PV problems, this technique is very useful.



**Figure 1. The process of transcribing Chinese text into Taiwanese pronunciation using the ASR technique.**

In this paper, we report two experiments using speech and text data, called the Taiwanese Buddhist Sutra (TBS) corpus [Sik 2004b]. The phonetic transcription framework is described in Section 2. Given a speech corpus with phonetic transcription for training, Section 3 reports the speech recognition results with and without the corresponding text for its phonetic transcription. Section 4 discusses the second experiment involving speech recognition with the corresponding text under various pronunciation variation conditions in the training corpus. Section 5 presents our conclusions.

## 2. The Phonetic Transcription Augmented by Speech Recognition Technique

Figure 2 is the framework of phonetic transcription using the speech recognition technique. While the input is a speech waveform and Chinese Sutra text, the output is a phonetic transcription corresponding to the input Chinese text. The entire framework can be divided into two major parts, *i.e.* an acoustic part and a linguistic part.

Based on flow chart in Figure 2, we define the following notations:  $\underline{s}$  is the syllable sequence, while  $\underline{c}$  and  $\underline{o}$  are the input character and augmented acoustic sequences. The phonetic transcription target is to find the most probable syllable sequence  $\underline{s}^*$  given  $\underline{o}$  and  $\underline{c}$ . The formula is:

$$\underline{s}^* = \arg \max_{\forall \underline{s} \in \underline{S}} P(\underline{s} | \underline{o}, \underline{c}) \quad (1)$$

Where  $\underline{c} \in \underline{C} = \{ \underline{c} | \underline{c} = c_1^M = c_1 \dots c_M, c_i \in C \}$ ,  $c_i$  is an arbitrary Chinese character,  $C$  is the set of all Chinese characters, and the number of elements in  $C$  is  $n(C) \approx 13000$ .  $\underline{s} \in \underline{S} = \{ \underline{s} | \underline{s} = s_1^N = s_1 \dots s_N, s_i \in S \}$ ,  $s_i$  is an arbitrary Taiwanese syllables,  $S$  is the set of all Taiwanese syllables, and the number of elements in  $S$  is  $n(S) \approx 1000$ . Using the Bayes theorem:

$$\underline{s}^* = \arg \max_{\forall \underline{s} \in \underline{S}} \frac{P(\underline{s} | \underline{c}) P(\underline{o} | \underline{s}, \underline{c})}{P(\underline{o} | \underline{c})} \quad (2)$$

The acoustic sequence  $\underline{o}$  is assumed dependent only on the syllable sequence  $\underline{s}$ . Equation 2 could be simplified as:

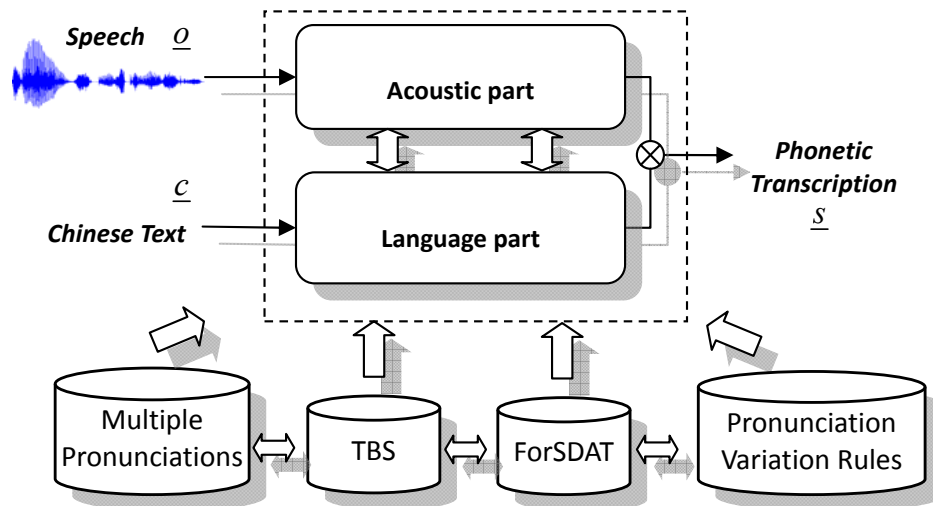
$$\underline{s}^* = \arg \max_{\forall \underline{s} \in \underline{S}} P(\underline{s} | \underline{c}) P(\underline{o} | \underline{s}) \quad (3)$$

The first term,  $P(\underline{s} | \underline{c})$ , of Equation 3 is independent of  $\underline{o}$  and plays the major role in the linguistic part of the recognition scheme. The second term,  $P(\underline{o} | \underline{s})$ , is the probability of observation given the syllable sequence, which plays the major role in the acoustic part.

For the acoustic part, which is the probability of observing an acoustic sequence  $\underline{o}$ , given a phonetic syllable sequence  $\underline{s}$ , it is well known that the Hidden Markov Model (HMM) can be used to model it. We can choose a speaker independent HMM model (SI-HMM) with



speaker adaptation techniques.



**Figure 2. The flow chart of the phonetic transcription of Taiwanese Buddhist Sutra (TBS) incorporating pronunciation variation rules.**

The linguistic part, which is the probability of observing a syllable sequence  $\underline{s}$ , given a character sequence  $\underline{c}$ , could be modeled as a traditional grapheme-to-phoneme problem. In such a problem, a “well-coverage” phonetic lexicon, which covers as many as possible correct pronunciations for each phoneme, is quite useful. The problem of multiple pronunciations could be solved using a specially designed searching net, such as the sausage net, which was named for its shape being similar to a sausage. All the searching nets, including the sausage net, were constructed according to a multiple pronunciation lexicon and described in the next section. Even the best pronunciation lexicon would miss the true pronunciation for a certain Chinese character. This is severe, especially for a minority language without many linguistic resources, like Taiwanese. To address this issue, the pronunciation variation rules would be incorporated in a sausage net to improve the accuracy of transcription.

### 3. Solutions to Multiple Pronunciation Problem

The first experiment is performed on the Sutra phonetic transcription using the sausage recognition network without considering the pronunciation variation problem. For a syllabic language such as Taiwanese or Mandarin, we can construct a concatenated net of all syllables. Based on Equation 3, we define:  $\underline{s} = s_1, s_2, \dots, s_N$  as the syllable sequence. As our goal is to find the real pronunciation, it would not be crucial to know the relationship between Chinese characters and syllables. Therefore, assume that the underlined character sequence  $\underline{c}$  is known and independent of  $\underline{s}$ , and all syllables are independent of each other. Following Equation 3, we have:

$$\begin{aligned}
\underline{s}^* &= \arg \max_{\forall \underline{s} \in \underline{S}} P(\underline{s})P(\underline{o} | \underline{s}) \\
&= \arg \max_{\forall \underline{s} \in \underline{S}} P(s_1)P(s_2)\dots P(s_N)P(\underline{o} | s_1, s_2, \dots, s_N) \\
&= \arg \max_{\forall \underline{s} \in \underline{S}} P(\underline{o} | s_1, s_2, \dots, s_N) \prod_{i=1}^N P(s_i)
\end{aligned} \tag{4}$$

To make the case simple and straightforward, we could assume that  $P(s_i)$  is a uniform distribution, then Equation 4 can be simplified as follows:

$$\underline{s}^* = \arg \max_{\forall \underline{s} \in \underline{S}} \left( \frac{1}{n(S)} \right)^N P(\underline{o} | s_1, s_2, \dots, s_N) = \arg \max_{\forall \underline{s} \in \underline{S}} P(\underline{o} | s_1, s_2, \dots, s_N) \tag{5}$$

where  $S$  is the set of all possible syllables and the  $n(S)$  is the number of elements in  $S$ .

Such a concatenated net is called a total-syllable net. It is a compact representation of the searching space  $\underline{S}$ , which is a set of all possible syllable sequences. The transcription performance in this way is dependent only on the acoustic part. Therefore, the experimental results conducted using a total-syllable net is referred to as the performance of the acoustic part.

Second, it is also possible to perform the phonetic transcription using only text input without any speech/acoustic clues. This is the linguistic part in the recognition scheme shown in Fig. 2. In this case, Equation 3 can be simplified as:

$$\underline{s}^* = \arg \max_{\forall \underline{s} \in \underline{S}} P_{\underline{S}|\underline{C}}(\underline{s} | \underline{c}) = P_{S_1 \dots S_N | C_1 \dots C_N}(s_1 \dots s_i \dots s_N | c_1 \dots c_i \dots c_N) \tag{6}$$

As only a small scale database is available, we assume that  $s_i$  is dependent on  $c_i$  and  $s_{i-1}$ , or even only on  $c_i$ . Equation 6 can then be simplified as:

$$\underline{s}^* = \arg \max_{\forall s_i \in S_i} \prod_{i=1}^N P_{S_i | C_i}(s_i | c_i) \tag{7}$$

and

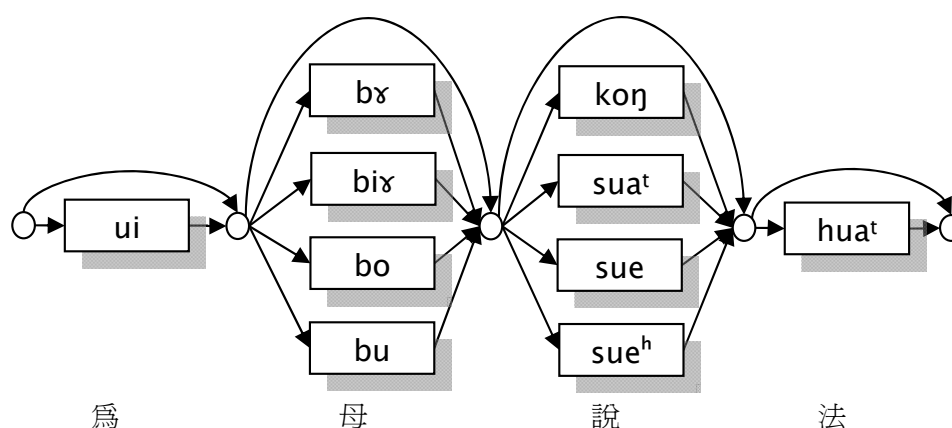
$$\underline{s}^* = \arg \max_{\forall s_i \in S_i} \prod_{i=1}^N P_{S_i | C_i}(s_i | s_{i-1}, c_i) \tag{8}$$

The results from the experiments conducted using Equations. 7 and 8 depend only on the textual input instead of the acoustic input, and are referred to as the language part performance. Therefore, discussion about Equations 5 and 7 require traditional automatic speech recognition and grapheme-to-phoneme approaches for dealing with the phonetic transcription tasks.

What is proposed in this paper is an approach to integrate both. Given a Chinese character sequence, based on the multiple pronunciations of each Chinese character, a much smaller recognition net can be constructed. Thus, by integrating Equations 5 and 7, we have:

$$\underline{s}^* = \arg \max_{\forall s_i \in \mathcal{S}_i} P(\underline{s} | s_1, s_2, \dots, s_N) \prod_{i=1}^N P(s_i) P_{\mathcal{S}_i | C_i}(s_i | c_i) \quad (9)$$

Taking an example of a typical text sentence “爲母說法”, which is shown in Figure 3, we will call such a net (with multiple pronunciations) a sausage net. Higher recognition accuracy can be expected due to the smaller perplexity in the recognition net. Our task is to construct “good” sausage nets to help the acoustic part do the job. In the following, we will discuss how to use the lexicons, the recognition networks to implement the proposed framework and show some experiment results.



**Figure 3.** The sausage searching net. The net is constructed from the multiple pronunciations of each Chinese character from the Formosa Lexicon. The corresponding Chinese characters with multiple pronunciations are also shown.

### 3.1 Speech Database

In this paper, we use as the speech database, Formosa Speech Database (ForSDAT), which was collected over the past several years [Lyu *et al.* 2004]. The SI-HMM model can be trained from the ForSDAT-01, which contains 200 speakers and 23 hours of speech. All speech data were recorded in 16K, 16bit PCM format. The statistical information of ForsDAT-01 was summarized as in Table 2.

In addition, the partial ForSDAT-02 speech corpus was used to derive the rule set of pronunciation variations, which contains 131 speakers and 7.2 hours of speech. The statistical information of partial ForsDAT-02 was summarized as in Table 2 and the detail is discussed in Section 4.

The distribution of another speech database, TBS, is listed in Table 3, where there are 1,619 utterances in this speech data set with a total length of about 230 minutes [Sik 2004b]. 502 utterances, which include 5909 syllables, are randomly chosen and reserved for testing

while as another 31 utterances are used for acoustic model development.

**Table 2. The statistics of ForSDAT-01 speech corpus and partially manually validated ForSDAT-02 speech corpus.**

	<b>ForSDAT-01</b>	<b>Partial ForSDAT-02</b>
<b>Utterance</b>	92158	19731
<b>Number of People</b>	100 (male: 50, female: 50)	131 (male: 72, female: 59)
<b>Number of Syllable</b>	179730	45865
<b>Number of Distinct Triphones</b>	1356	1194
<b>Number of Total Triphones</b>	555731	104894
<b>Time (hr)</b>	22.43	7.2

**Table 3. TBS (Taiwanese Buddhist Sutra) speech corpus.**

<b>Buddhist Corpus Category</b>	<b>Utterance</b>	<b># of Syllable</b>	<b>Time (min)</b>
<b>Adaptation</b>	31	359	2.62
<b>Test</b>	502	5909	43.23
<b>Other</b>	1086	12147	179.88
<b>Total</b>	<b>1619</b>	<b>18415</b>	<b>225.73</b>

### 3.2 The Pronunciation Lexica

There is one pronunciation lexicon available to us, **Formosa Lexicon**, which provides multiple pronunciations in Taiwanese for all Chinese characters. The lexicon contains about 123,000 words in Chinese/Taiwanese text with Mandarin/Taiwanese pronunciations. It is a combination of two lexica: Formosa Mandarin-Taiwanese Bi-lingual lexicon and Gang's Taiwanese lexicon [Liang *et al.* 2004a; Liang *et al.* 2004b; Lyu *et al.* 2004]. The former is derived from a Mandarin lexicon; thus, many commonly used Taiwanese terms are missing due to the fundamental difference between these two languages. The latter contains more widely used Taiwanese expressions from samples of radio talk shows. Some examples of the lexicon are shown in Table 4, containing 65,007 entries using the Wen-du-in pronunciation and 58,431 entries using the Bai-du-in pronunciation. There are a total of 123,438 pronunciation entries. For all 65,007 Wen-du-in pronunciation entries, there are 6,890 entries for one-syllable words, 39,840 entries for two-syllable words, and so on. The lexicon as described above is a general-purpose lexicon. It could be used for a wide range of applications and tends to have a higher number of multiple pronunciations.

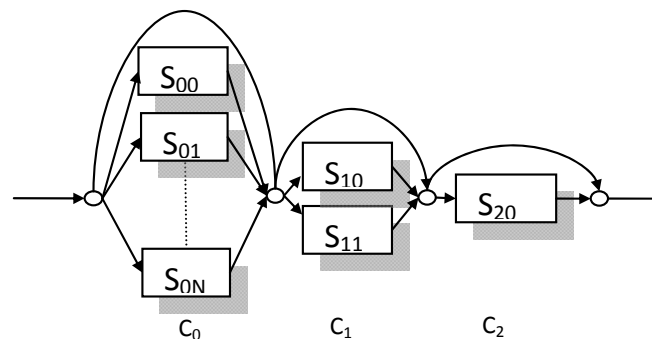
We used two kinds of searching nets in these experiments, according to the lexicon. The first is the Total-syllable net. It is simply a concatenated net of all Taiwanese syllables existing in the Taiwanese Buddhist Sutra (TBS), where the total number of syllables is 467, denoted as the Total-Syl-Net. The other searching nets are the sausage nets generated from

each of the pronunciation lexica. The nets were constructed by filling in each node of the net with the corresponding multiple pronunciations of each Chinese character from the pronunciation lexicon. One example is shown in Figure 3. The nets are denoted the General-Sau-Net for the general-purpose Formosa Lexicon.

**Table 4. The partial example of all possible pronunciations per Chinese Character from Formosa bi-lingual Lexicons, including classic literature pronunciation (Wen-du-in) or daily life pronunciation (Bai-du-in).**

	Pronunciation 1	Pronunciation 2	Pronunciation 3	.....
日(sun)	gí <sup>p</sup>	lí <sup>p</sup>	zǐ <sup>p</sup>	.....
火(fire)	xê	xùe		
加(add)	ká	ké	kúe	.....
叩(knock)	k'áu	kìx	k'ò <sup>k</sup>	.....
卵(egg)	lĭj	lûan	nĭj	.....
坐(sit)	tsé	tsǎ	tsūe	.....

However, a lexicon is inevitably incomplete, and we could be confronted with the missing character problem and the missing pronunciation problem. The missing character problem is when a character used in the Sutra does not appear in the lexicon. One reason is because many of the Chinese characters used in ancient times are no longer used in modern times. Thus, even the Unicode Standard, which contains more than thirty thousand Chinese characters, does not contain them. The Formosa Lexicon has much fewer distinct characters, and the missing character problem is inevitable. When a missing character is encountered, we use all possible syllables as its multiple pronunciations. One example is illustrated in Figure 4, where the sausage searching net is constructed for the Chinese character string “ $C_0C_1C_2$ ”. It is assumed that the character  $C_0$  is a missing character. In such a case, all possible syllables, denoted as  $S_{00}, S_{01}, \dots, S_{0N}$ , are used as possible pronunciations of  $C_0$ .



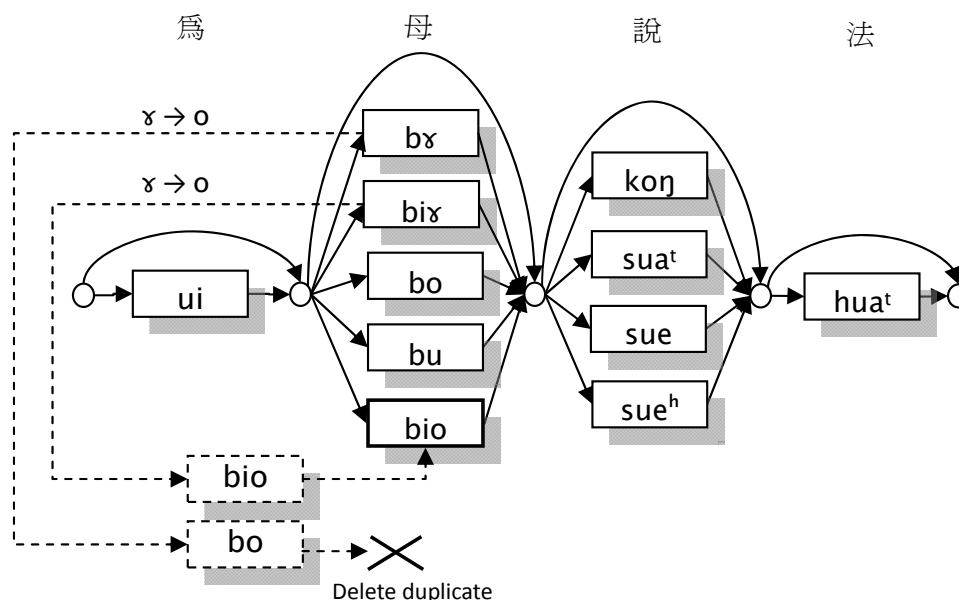
**Figure 4. The sausage searching net with missing character  $C_0$ , where all syllables,  $S_{00}, S_{01}, \dots, S_{0N}$ , are used as its possible pronunciations.**

#### 4. Incorporating Pronunciation Variation Rules

As insufficient coverage of pronunciation nodes in the searching net will severely degrade the recognition performance, some approaches to extend the pronunciation coverage will be considered to help the overall performance. Since global lexicon modification by experts would take considerable effort and not necessarily benefit, we adopted alternative rule-based methods. By rule-based pronunciation variations, we mean that phonetic units will be changed by speakers according to some underlying rules. Usually, a rule could be notated as the form “ $B \rightarrow S$ ” for canonical pronunciation  $B$  (base-form) being substituted with the actual pronunciation  $S$  (surface-form) [Saraclar *et al.* 2004]. Briefly speaking, some rule-derived variant pronunciations are added directly into the searching net to enhance the poor pronunciation coverage of an imperfect pronunciation lexicon.

An example is shown in Figure 5, where the number of pronunciations for the Chinese character “母” was increased from 4 to 5 by incorporating some specific pronunciation rules as “ $/\gamma/ \rightarrow /o/$ ”. It could be shown that, as long as the pronunciation rules could be well designed, the phonetic transcription performance would be effectively improved.

Generally speaking, the pronunciation-variation (PV) rules can be categorized into two kinds: knowledge-based and data-driven rules. The knowledge-based rules were derived from the knowledge established by phoneticians. On the other hand, the data-driven PV rules rely on the availability of transcribed speech corpora.



**Figure 5.** An example of the extended sausage searching net. The net is constructed from the multiple pronunciations in lexicon and expanded using pronunciation-variation rules for each Chinese character according to the rule “ $/\gamma/ \rightarrow /o/$ ”.

### 4.1 The Knowledge-Based Variation Rules

Considering the trade-off between the number of elements and the degree of detail or perplexity, the triphone was used as the acoustic unit, thus, the transcription unit in this paper. The form  $LBR \rightarrow LSR$  represents the pronunciation variation rules, where  $B$  and  $S$  represent the base form and surface form of a central phone, and  $L, R$  are the left and right contexts respectively. The number of triphone units in Taiwanese is about 1200.

As with the other members of the Chinese language family, there are about three types of pronunciation variations in Taiwanese. These could be summarized as follows and are shown in Table 5:

1. Variation between Bai-du-in and Wen-du-in: The variations may vary due to using classic literate pronunciation (known as Wen-du-in) or a colloquial pronunciation (known as Bai-du-in). This point has been discussed previously in Section 1. For example, the Chinese character “生” (to give birth) might be pronounced as /siŋ/ in Wen-du-in and /sẽ/ or / sī / in Bai-du-in. All of these are acceptable.
2. Variation between sub-dialectal regions: Some variations were referred to as dialectal differences; For instance, the initials /z/ is substituted with /l/ or /g/ depending on the sub-dialect of Taiwanese. Such a rule was denoted as “z→l/g”.
3. Variation due to personal pronunciation errors: Some kinds of variations are considered personal pronunciation errors. Owing to the lack of some phonemes in the mainstream language (such as Mandarin in Taiwan), some pronunciation may disappear in younger generations. One of these phonemes is /g/, where the phenomenon is denoted as “g → {}”.

Table 5 can be considered as a knowledge source to select the pronunciation variation rules. The knowledge-based PV rules, which were derived by more than one linguist, were sometimes contradictory with each other. This made them difficult to choose between at times in implementation. Such a difficulty leads to a need for another approach. One of them is a data-driven approach from large-scale real data. Since a little manually transcribed speech data was available, we could use statistical computational measures to extract the PV rules from real data. This issue will be discussed in the following section.

**Table 5. The three types of pronunciation variations in Taiwanese.**

Variation types	Examples
Bai-du-in / Wen-du-in	siŋ → sẽ/sī
Dialectal difference	z → l/g ɣ ↔ ɔ → o n → l b → m ĩũ → ãõ
Personal pronunciation error	g → {} b → {} h → {}

## 4.2 The Sata-Driven Variation Rules

The same simple way to adopt the methodology of pronunciation variation is to expand the pronunciation lexicon using variation rules of the form  $LBR \rightarrow LSR$ . Similar work for such an approach was shown in Mandarin [Tsai *et al.* 2002]. To derive such rules, a speech corpus with both canonical pronunciation and actual pronunciation is necessary. We choose a subset of ForSDAT, called ForSDAT-02, to derive PV rules, and the statistical information is summarized as in Table 2.

ForSDAT-02 is a speech database with rich bi-phone coverage. This database was recorded by prompting speakers with a script. Although the script in Taiwanese text was shown with phonetic transcription, we did observe variations in the recorded speech. A small portion of the speech data was then manually checked, and the phonetic transcription of the script was corrected according to actual speech. Some examples of the original transcription (the base-form) and the manually corrected transcription (the surface-form) are shown in Table 6, which is called the sentence-level confusion table.

**Table 6. Sentence-level confusion table. The output is manually corrected transcription (the surface-form), and the input is the original transcription (the base-form).**

Original transcription (base-form)	Manually corrected transcription (surface-form)
è bɿ k'î ǎ	è bǒ k'î ǎ
gám k'ài bàn ts'én	gán k'ài bàn ts'én
sĩŋ ùa <sup>h</sup> hù <sup>h</sup> hǔa <sup>h</sup>	sĩn ùa <sup>h</sup> hù <sup>h</sup> hǔa <sup>h</sup>
.....	.....

From the sentence-level confusion table, it is quite a straightforward process to construct other confusion tables in syllable level and triphone level. These two tables are shown in Table 7 and Table 8 as follows.

**Table 7. Syllable-level confusion table, where  $z_{ij}$  represents the number of variation from syllable  $x_i$  (base-form) to triphone  $y_j$  (surface-form),  $T$  is the number of surface-form and base-form**

	bɿ	bo	.....	$y_j$	.....	sĩŋ	sin
bɿ	237	30	.....	$z_{1j}$	.....	0	0
bo	0	64	.....	$z_{2j}$	.....	0	0
.....	...	...	.....	...	.....		
$x_i$	$z_{i1}$	$z_{i2}$	.....	$z_{ij}$	.....	$z_{i,T-1}$	$z_{iT}$
.....	...	...	.....	...	.....		
sĩŋ	0	0	.....	$z_{T-1,j}$	.....	163	12
sin	0	0	.....	$z_{T,j}$	.....	2	105



The triphone-level confusion table is used as a direct knowledge source to derive the PV rules, where each cell in the table was looked upon as a rule. The number of rules shown in Table 8 is  $P^2$ , where  $P$  is the number of triphones (about 1200 in the target language). The number of rule set selections is  $2^{P^2}$ , which is an enormous number which is impossible to be processed in modern computers. To make the problem more solvable, some specially designed algorithms should be developed that are able to specifically find a useful route in the huge rule set selection space within a reasonable time.

**Table 8. Triphone-level confusion table, where  $n_{ij}$  represents the number of variation from triphone  $b_i$  to triphone  $s_j$ ,  $P$  is the number of surface-form and base-form,  $N_i = \sum_j n_{ij}$ ,  $M_j = \sum_i n_{ij}$  and  $N = \sum_i \sum_j n_{ij}$**

	<i>bh-er</i>	<i>i-ng</i>	<i>i-n</i>	.....	$s_j$	.....	<i>bh-o</i>	<i>a-n</i>	<i>a-m</i>	
<i>bh-er</i>	237	0	0	.....	$n_{1j}$	.....	30	0	0	267
<i>i-ng</i>	0	1273	84	.....	$n_{2j}$	.....	0	0	0	1373
...	...	...	...	.....	...	.....	...	...	...	
$b_i$	$n_{i1}$	$n_{i2}$	$n_{i3}$	.....	$n_{ij}$	.....	$n_{i,p-2}$	$n_{i,p-1}$	$n_{iP}$	$N_i$
...	...	...	...	.....	.....	.....	...	...	...	
<i>a-m</i>	0	0	0	.....	$n_{pj}$	.....	0	35	834	
	241	1315	1102	.....	$M_j$	.....	107	1873	870	$N$

First of all, some criteria should be adopted to choose the most significant rule sets. Three kinds of statistical measures were used in this paper. They are (1) Joint probability [Raux 2004], (2) Conditional probability, and (3) Mutual information-like of the base form pronunciation and the surface form pronunciation. The mathematical definitions of the above three measures are as follows:

1. Joint probability of the base form pronunciation  $b_i$ , and the surface form pronunciation  $s_j$ ,

$$p(b_i, s_j) = n_{ij} / N$$

2. Conditional probability of the surface form pronunciation  $s_j$ , conditioned on the base form pronunciation  $b_i$ ,

$$p(s_j | b_i) = n_{ij} / N_i$$

3. Mutual information of the base form pronunciation  $b_i$ , and the surface form pronunciation  $s_j$ ,

$$I_{ij} = p(b_i, s_j) \log \frac{p(b_i, s_j)}{p(b_i)p(s_j)} = \frac{n_{ij}}{N} \log \left( N \frac{n_{ij}}{\sum_i n_{ij} \cdot \sum_j n_{ij}} \right)$$

In all of the above equations,  $n_{ij}$  is the number of (base-form) triphone  $b_i$  substitutions by the surface-form triphone  $s_j$  that appear in a corpus, and

$$N = \sum_i \sum_j n_{ij},$$

$$N_i = \sum_j n_{ij},$$

$p(b_i, s_j)$  represents the joint probability of  $(b_i, s_j)$ ,

$p(b_i)$  and  $p(s_j)$  equal the marginal probability of  $b_i$  and  $s_j$ , respectively.

Note that each pair  $(i, j), i \neq j$ , corresponds to a substitution rule, and we select those pairs  $(i, j)$  with higher scores of  $p(b_i, s_j)$ ,  $p(b_i, s_j)$  and  $I_{ij}$  to be the variation rules to extend the sausage net pronunciation.

In Table 9, the rules were sorted by rank based on joint probability, conditional-probability, and mutual-information. There are variants among the three lists. One rule which is much more important in some method may be trivial in the other method.

**Table 9. Data-driven rules: The top 10 substitution errors were listed from the partially validated ForSDAT-02 corpus for Joint-Probability-Based, Conditional-Probability-Based and Mutual-information-Based method**

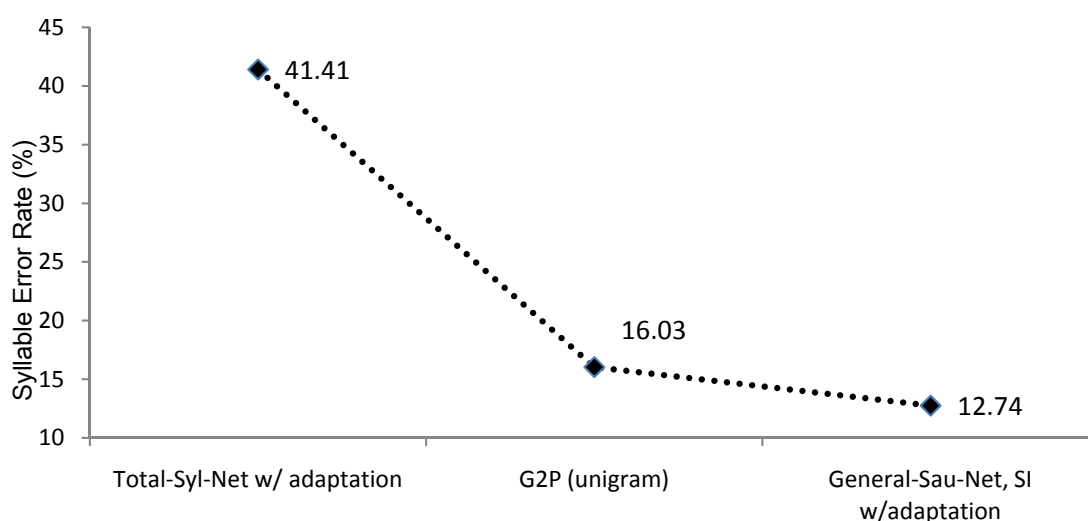
Rank based on Joint Probability	Rank based on Conditional Probability	Rank based on Mutual-Information
i-ŋ → i-n	x-ā <sup>h</sup> → x-a <sup>h</sup>	i-ŋ → i-n
a-m → a-n	<b>n-ʒ</b> → <b>n-ō</b>	<b>b-ʒ</b> → <b>b-o</b>
<b>b-ʒ</b> → <b>b-o</b>	ŋ-ĩ → ŋ-ě	ĩ-ō → ĩ-ũ
i-m → i-n	l-o <sup>h</sup> → l-o	l-i <sup>k</sup> → l-i <sup>t</sup>
ĩ-ō → ĩ-ũ	k'-i <sup>h</sup> → k'-i <sup>k</sup>	<b>k'-ʒ</b> → <b>k'-o</b>
a-n → a-m	ts-a <sup>k</sup> → ts-a <sup>t</sup>	ĩ-ŋ → ĩ-n
a-ŋ → a-m	<b>g-ʒ</b> → <b>g-o</b>	<b>p-ʒ</b> → <b>p-o</b>
i-a-ŋ → i-o-ŋ	p'-i <sup>k</sup> → p'-i <sup>t</sup>	i-m → i-n
i-m → i-ŋ	ĩ-ō → ĩ-ũ	<b>t-ʒ</b> → <b>t-o</b>
i-n → i-m	g-i <sup>k</sup> → g-i	b-i <sup>t</sup> → b-i <sup>h</sup>

## 5. The Experiment Results and Discussion

In training or estimating SI-HMM models for the acoustic part, we use continuous Gaussian-mixture HMM models with feature vectors of 12-dimensional MFCC with 1-dimensional energy, plus the first, second, and third derivatives computed using a 20-ms frame width and 10-ms frame shift. Context-dependent intra-syllabic tri-phone models were built using a decision-tree state tying procedure. As the testing data is speaker dependent, adaptation with some manually transcribed data must be useful in automatic phonetic

transcription. Maximum Likelihood Linear Regression (MLLR) is then used to adapt speaker independent models using 31-utterance adaptation speech data. Most of the training and recognition are carried out by using the HTK tools [Young *et al.* 2008].

With the two searching nets (Total-Syl-Net, General-Sau-Net) and acoustic models (SI with adaptation), the recognition results measured as the syllable error rate (SER) are shown in Figure 6. In addition, we also show the result of only language, called grapheme-to-phoneme (G2P), with unigram.



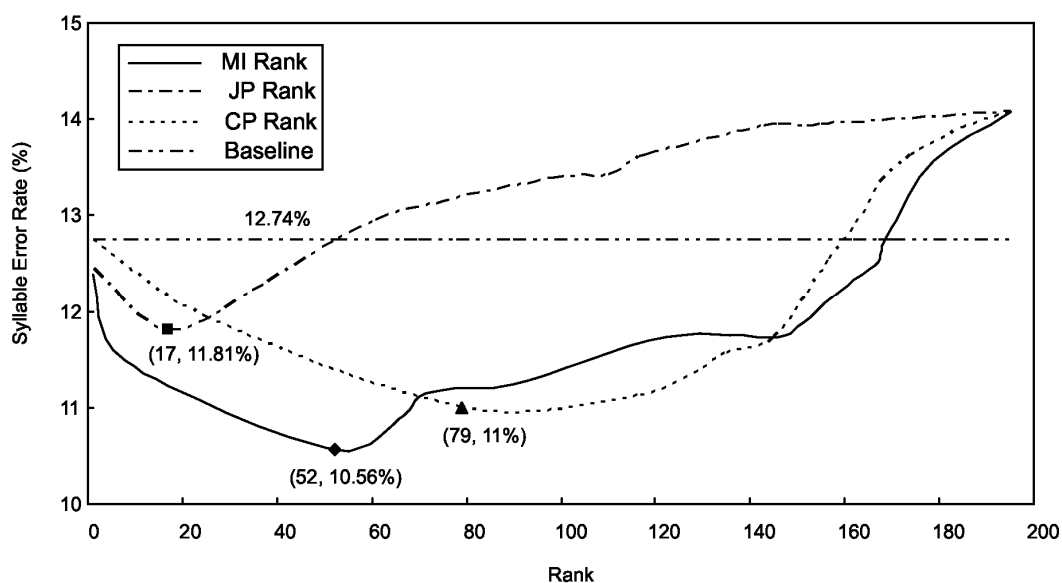
**Figure 6. Syllable error rate (SER) under uni-gram, and General-Sau-Net, SI w/ adaptation. See text in subsection 3.2 for notations.**

Through observation of the experimental results, we can see that neither G2P with unigram nor Total-Syl-Net with adaptation model can reach acceptable performance. Therefore, it is necessary to integrate the linguistic and acoustic parts. General-Sau-Net could surpass Total-Syl-Net. For example, under the same speaker adaptation models, the result was 12.74% better with General-Sau-Net than other results in Figure 6. Thus, if the speaker independent model could be adapted using some phonetically transcribed speech data, the adapted speaker independent model under General-Sau-Net would be suitable for the phonetic annotation task. In multiple pronunciation problems, by our speech data observation, we could see that some errors result from pronunciation variations. Therefore, we hypothesized that the performance would get better by adaptation of the Formosa Lexicon Sausage net, *i.e.* adaptation of the Formosa lexicon, described in the Section 4.

We adapted the Formosa (general-purpose) pronunciation lexicon according to different pronunciation variation rule sets. The speech recognition task with a sausage searching net and speaker adapted acoustic models was then conducted, as described in Section 4, wherein, the

SER achieved before the application of the pronunciation variation rules was **12.74%**, as shown in Figure 6. This would be looked upon as the performance of the baseline setup in this section.

In Figure 7, the transcription performance was measured in terms of syllable error rate vs. the number of ranked PV rules sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the baseline setup. We could observe that it is truly helpful to decrease the SER by increasing the searching net coverage via the PV rules. The evidence is that the lowest error rate (**10.56%**) was achieved by utilizing the first 52 variation rules, which were selected by the Mutual-Information (MI) measure. Similar improvement would also be observed in the best SER (**11.81%** and **11%**) achieved using the Joint-Probability (JP) and Conditional-Probability (CP) measures when the complexities of the JP and CP measure are 2.68 and 2.55, respectively.



**Figure 7. The recognition result (Syllable error rate) v.s. the number of ranked rules sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion.**

It is interesting to point out that, in Figure 7, choosing different statistical measures was found to influence the achievable lowest SER and also the speed of decrease in SER. In these experiments, we found that MI was the best in terms of the rate of decrease in SER or the achievable lowest SER. Although the JP-based measure could make the error rate converge more quickly than the CP-based measure, the performance also degraded quickly. This is because the CP-based measure score was normalized by the base-form count in contrast to the

JP-based measure. However, the insignificant and harmless PV-rules might get the higher conditional probability sometimes due to few base-form observations. The PV-rules for the CP-based measure might not increase the perplexity but still lead to the slowest convergence among the three measures. In the MI-based measure, the formula could avoid slow convergence using the Joint-Probability as a weight when the base-form had few variations. Observing the confusion table, the surface-form would have lower correlation with these base-forms if many base-forms would transform into the same surface-form. So, we proposed that the mutual information between the base and surface-forms should be used to calculate the base and surface-form correlation using the normalization of their count. Consequently, the error rate of the MI rank converges most quickly and the performance of the MI measure in error reduction was also better than JP and CP measures, respectively.

Another interesting point was that the SER will possibly increase if too many PV rules are applied. For example, the lowest SER is achieved by applying 52 rules when MI was adopted as the ranking measure. However, after applying more rules, the SER increased! It even became worse than that in the baseline experiments. This means that some “bad” pronunciation variation rules may lead to a performance reduction. Take the Joint-Probability (JP) measure for example. The optimal performance was achieved when 17 ranked rules were applied, but when the number of rules further increased, the performance degraded. It was similar when MI or CP was used. Therefore, it is important to determine “good” rules and choose them so that the optimal performance could be achieved as soon as possible.

Extending the searching net can enhance the SER performance, but the extension must be limited to a suitable range. This point can be observed from the perplexity of the searching net in Figure 8. Regardless what measures we use, the differences in the perplexity values from the best results among the three measures were always slight. For example, in Figure 8, the perplexity of the best JP measure result was 2.68 when the perplexity of best MI measure result was 2.62. That means too many rules may lead to more real pronunciation coverage, but the performance may improve slightly or even decrease progressively. The perplexity is a good measure to evaluate the searching net in obtaining the best results.

Finally, in Figure 9, the error rate of General-Sau-Net is 12.74%. However, some errors resulted from pronunciation variations caused by a speaker's accent. Therefore, through incorporating variation rules into General-Sau-Net with different statistic measures, the best error rates can be reduced to 11.81%, 11%, and 10.56% with respect to JP-, CP-, and MI-based measures, respectively.

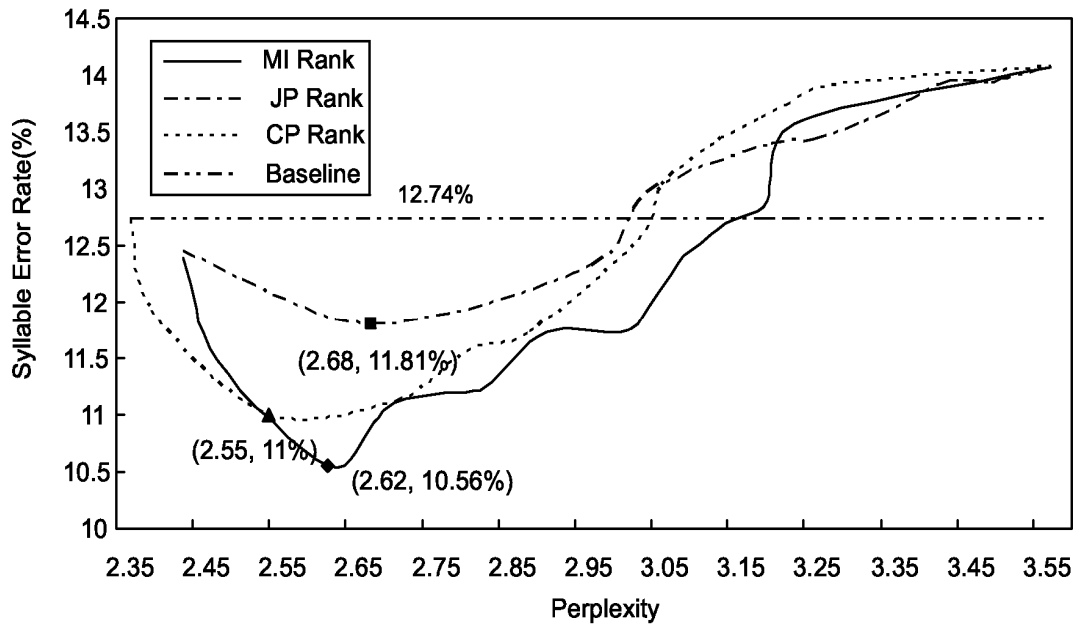


Figure 8. The recognition result (syllable error rate) vs. the perplexity sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion

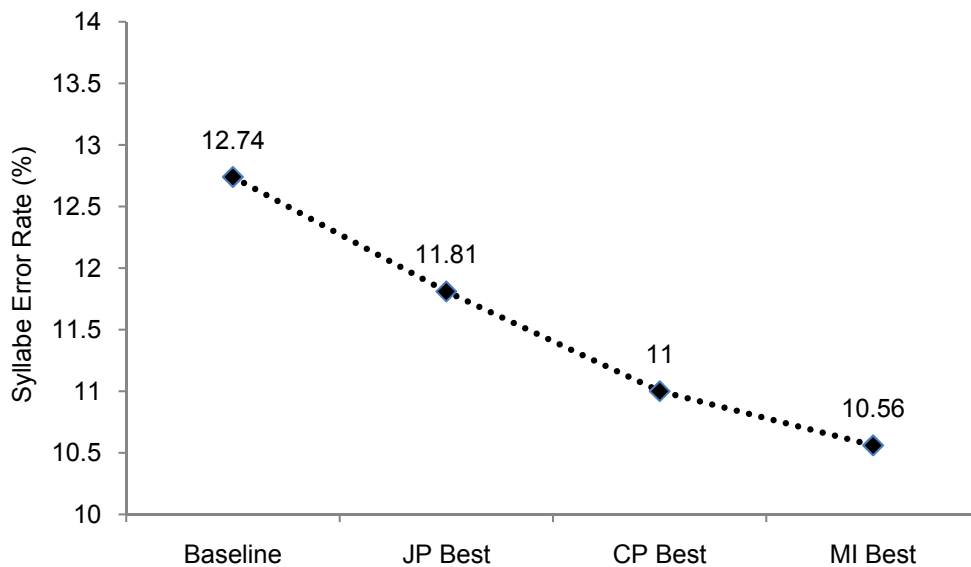


Figure 9. The recognition result (syllable error rate) under four kinds of net according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion.

## 6. Conclusions

We have proposed a new approach to address the phonetic transcription of Chinese text into Taiwanese pronunciation. Considering the fact that there are very few linguistic resources for Taiwanese, we used speech recognition techniques to deal with multiple pronunciation variations, which is a very common phenomenon in Taiwanese but hard to deal with using traditional text-based approaches. A general-purpose lexicon (called the Formosa lexicon), and a speaker-adapted HMM model were used to achieve a syllable error rate, 12.74%. In order to enhance the performance, the trivial adaptation of a general-purpose sausage net with pronunciation variation rules was used instead of global pronunciation lexicon modification.

In pronunciation variation rule (PV-rules) selection, the data-driven variation rules, which were derived using three statistical measures, were used to extend more possible pronunciations. Although the knowledge-based rules were also derived from a knowledge source, the rules were difficult to implement and dependent on the specific language. Thus, we selected data-driven rules with context-dependent triphones as the general solution to the PV problem. In the data-driven measure, the mutual-information-based (MI) rank outperformed the Joint-Probability rank and Conditional-Probability rank. Compared with baseline experiment result error rate of 12.74%, the lowest error rate of the MI-based measures had an error reduction rate of 17.11%, which was the best among the three statistical measures proposed in this paper. The error rate of the MI rank converged most quickly and the best performance of MI-based measure appeared in the first 52 ranks.

The experimental results from data-driven measures could possibly provide the evidence to help choose the corresponding knowledge-based PV rules. Of course, some of the pronunciation variation rules were certainly language-dependent (*i.e.* the phonological and phonetic processes differ between languages) [Kanokphara *et al.* 2003]. However, the major points to be emphasized are that the proposed technique to model pronunciation variation for transcription was rather language-independent.

The recognition of tones was still an unsolved problem in this research. This is another issue for further research. In Taiwanese, there are 7 tone classes, which could be used to distinguish the meanings of words. In addition, the complex tone sandhi would also be accompanied with tone recognition. If more speech and text was gathered, the analysis and statistics of pronunciation information for pronunciation probability would be the next step. We will construct a human interaction system to help more Taiwanese publications be presented. This technology may also be used as a language-learning tool.

Although the proposed technique was developed for Taiwanese speech, it could also be easily adapted for application in other similar “minority” Chinese spoken languages, such as Hakka, Wu, Yue, Xiang, Gan, and Min, or other non-Han family languages which also use

Chinese characters as the written language form.

In summary, the proposed semi-automatic transcription of Chinese text into a Taiwanese pronunciation system reached a **12.74%** error rate in the baseline experiment. Further improvement using pronunciation variation rules produced a **17.11%** error rate reduction.

## Reference

- Chen C. H., Sutra on the original Vows of Bodhisattva Earth Treasure in English, <http://www.yogichen.org/efiles/b041a.html>, 2006.
- Cover, T. M. and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.
- Cremelie, N., and J.-P. Martens, "In Search of Better Pronunciation Models for Speech Recognition," *Speech Communication*, 29, 1999, pp. 115-136.
- Evermann, G., H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P.C. Woodland, "Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System," In: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, Montreal, Canada, pp. I-249-I-252.
- Haeb-Umbach, R., P. Beyerlein, and E. Thelen, "Automatic Transcription of Unknown Words in a Speech Recognition system," In: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, pp. 840-843.
- Hain, T., "Implicit modeling of pronunciation variation in automatic speech recognition," *Speech Communication*, 46, 2005, pp. 171-188.
- U.S. Department of State's Bureau of International Information Programs, IIP report, <http://usinfo.state.gov/>, Dec, 2003.
- Kanokphara, S., V. Tesprasit, and R. Thongprasirt, "Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database," In: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, Hong Kong, pp. I-764-I-767.
- Kim, D. Y., H.Y. Chan, G. Evermann, M.J.F. Gales, D. Mrva, K.C. Sim and P.C. Woodland, "Development of the CU-HTK 2004 Broadcast News Transcription Systems," In: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, Philadelphia, USA, pp. 861-864.
- Lamel, L., J.-L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, 16, 2002, pp. 115-129.
- Liang, M.-S., R.-C. Yang, Y.-C. Chiang, D.-C. Lyu and R.-Y. Lyu, "A Taiwanese Text-to-Speech System with Applications to Language Learning," In: *Proc. Int. Conf. on Advanced Learning Technologies (ICALT)*, 2004, Joensuu, Finland, pp. 91-95.
- Liang, M.-S., D.-C. Lyu, Y.-C. Chiang and R.-Y. Lyu, "Construct a Multi-Lingual Speech Corpus in Taiwan with Extracting Phonetically Balanced Articles," In: *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju Island, Korea.



- Lyu, R.Y., M.S. Liang, and Y.C. Chiang, "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin," *International Journal of Computational Linguistics & Chinese Language Processing (IJCLCLP)*, 9(2), August 2004, pp. 1-12.
- Nanjo, H., and T. Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 12, Jul. 2004, pp. 391-400.
- Nouza, J., D. Nejedlova, J. Zdansky and J. Kolorenc, "Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju, Korea.
- Raux, A., "Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju Island, Korea.
- Saraclar, M., and S. Khudanpur, "Pronunciation change in conversation speech and its implications for automatic speech recognition," *Computer Speech and Language*, 18, 2004, pp. 375-395.
- Sarada, G.L., and N. Hemalatha, T. Nagarajan and Hema A. Murthy, "Automatic Transcription of Continuous Speech using Unsupervised and Incremental Training," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju, Korea.
- Sik, D.-G., *The Four Basic Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.
- Sik, D.-G., *Earth Treasure Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.
- Siohan, O., B. Ramabhadran, and G. Zweig, "Speech Recognition Error Analysis on the English MALACH Corpus," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju Island, Korea.
- Soltau, H., B. Kingsbury, L. Mangu, D. Povey, G. Saon and G. Zweig, "The IBM 2004 Conversational Telephony System for Rich Transcription," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, Philadelphia, USA, pp. I-205-I-208.
- Tripitaka, S. S., Sutra on the original Vows of Bodhisattva Earth Treasure in Chinese. <http://book.bfn.org/article/0016.htm>, 2005.
- Tsai, M.Y., F.C. Chou, and L.S. Lee, "Improved pronunciation modeling by inverse word frequency and pronunciation entropy," *In: Proc. IEEE Int. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2002, pp. 53-56.
- Wu, J., and V. Gupta, "Application of Simultaneous Decoding Algorithm to Automatic Transcription of Known and Unknown Words," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, Phoenix, USA, pp. 589-592.
- Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. (Andrew) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book*, 3.4 ed., 2008.

