

Chinese Word Auto-Confirmation System using a Hybrid Approach

Jia-Lin Tsai, Cheng-Lung Sung and Wen-Lian Hsu
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,clsung,hsu}@iis.sinica.edu.tw

Abstract

In this paper, we present a Chinese word auto-confirmation (CWAC) system that uses a hybrid approach to take advantage of statistical and linguistic techniques. The task of our CWAC system is to auto-confirm whether an n -gram input ($n \geq 2$ and n -gram frequency ≥ 2) is a Chinese word.

The experimental results demonstrate that, for n -gram frequency ≥ 2 using OPENFIND searching results as the experimental large-scale corpus, the unknown word identification (UWI) performance of our CWAC system can achieve 96.95% precision, 86.12% recall and 91.21% F-measure. Comparing the performance of our CWAC system with a typical hybrid UWI system [e.g. Chen *et al.* 2002] based on the *Word Standard in Taiwan*, our CWAC system has an 8% increase in precision, and 18% increase in recall bearing in mind that such a comparison cannot possibly be all that fair since different corpora are used.

The CWAC system is the first one we have constructed. We will continue developing other CWAC systems and integrate them into a multi-CWAC system.

Keywords: natural language processing, word segmentation, unknown word

1. Introduction

For humans, word segmentation and word sense disambiguation (WSD) occur naturally when reading a sentence. These tasks are difficult for a computer as it is hard to give computers the ability to create unseen knowledge from texts. Unseen knowledge refers to contextual meaning, such as meaningful word-pairs [Tsai *et al.* 2002 and 2004] and unknown lexicon. Unknown lexicon identification identifies (1) unknown words (2) unknown word sense, (3) unknown parts-of-speech (POS) and (4) unknown word pronunciation. We found that, in the Academia Sinica Balanced Corpus (ASBC) 1.0 [CKIP 1995], about 60% of the words are not found in the Revised Mandarin Dictionary (重編國語辭典) [MPCM 1998], one most widely used Chinese

dictionary in Taiwan. Among these unknown words, 30% are high-frequency words (word frequency ≥ 3) and 33% belong to the name entity recognition (NER) category. Thus, auto-construction of unknown lexicons has become a critical component in developing many Chinese natural language processing (NLP) systems.

Unknown word identification (UWI), though essential, is still quite problematic in Chinese NLP. Summarized from [Chang *et al.* 1997; Lai *et al.* 2000; Chen *et al.* 2002; Sun *et al.* 2003; Zhang *et al.* 2003] and our observation, the difficulty of Chinese UWI is caused by the following observations:

1. As in other Asian languages, Chinese sentences are composed of strings of characters without delineating blank spaces that mark word boundaries.
2. Most of Chinese characters can either be used as a morpheme or a word. The Chinese character **花(flower)**, for example, can be either a free morpheme or part of a word, such as **花香**(fragrance of flowers) and **野花**(wild flowers).
3. Unknown words, which are usually compound words or proper names, are too numerous to list in a machine-readable dictionary (MRD).
4. In some cases, whether an n-gram is a word or not depends on the sentences. The tri-gram **名歌手**(famous singer) is a word in the sentence **他/是/名歌手** (He is a famous singer), but not a word in the sentence **他/是/無名/歌手** (He is an unknown singer). Note that the symbol / indicates word boundaries.

To resolve these issues, hybrid, statistical and linguistic approaches have been developed. In statistical approaches, researchers use common statistical features, such as maximum entropy [Chieu *et al.* 2002], association strength [Smadja 1993; Dunnin 1993], mutual information [Church 2000], ambiguous matching [Sproat *et al.* 1996], as well as multi-statistical features [Chang *et al.* 1997; Ma *et al.* 2003] for unknown word detection and extraction. In linguistic approaches, three major types of linguistic knowledge: morphology, syntax, and semantics, are used to identify unknown words. Recently, there is a trend toward hybrid approaches that take advantage of the merits of both statistical and linguistic approaches. Summarized from [Chang *et al.* 1997; Lai *et al.* 2000; Chen *et al.* 2002; Sun *et al.* 2003; Zhang *et al.* 2003], the merit of statistical approaches is their simplicity and efficiency, whereas the merit of linguistic approaches is their effectiveness in identifying low frequency unknown words.

Since there is a trade-off between recall and precision, deriving a hybrid approach that optimizes the combined F-measures has become a major focus in the UWI community [Chang *et al.* 1997; Chen *et al.* 2002]. To solve the precision-and-recall optimization, we introduce a Chinese word auto-confirmation (CWAC) system that can achieve high precision and recall. CWAC can automatically confirming whether an n-gram input is a Chinese word without human intervention.

This paper is structured as follows. In Section 2, we present a method for con-

structuring a CWAC system. Experimental results of the CWAC system are presented in Section 3. Conclusions and future directions are discussed in Section 4.

2. Development of the CWAC System

An n-gram extractor is developed to extract all n-grams with $n \geq 2$ and n-gram frequency ≥ 2 from test sentences to be the n-gram input for our CWAC system. Figure 1 is the flow chart of the CWAC system, which includes six processes and four supporting databases. 50,000 words were randomly selected from the CKIP lexicon (CKIP [1995]) to be the system dictionary. We adopt LFSL (linguistic approach first and statistical approach last) hybrid approach (see Figure 1) to develop our CWAC system. Specifically, processes 4 and 5 are linguistic approaches and process 6 is a statistical approach.

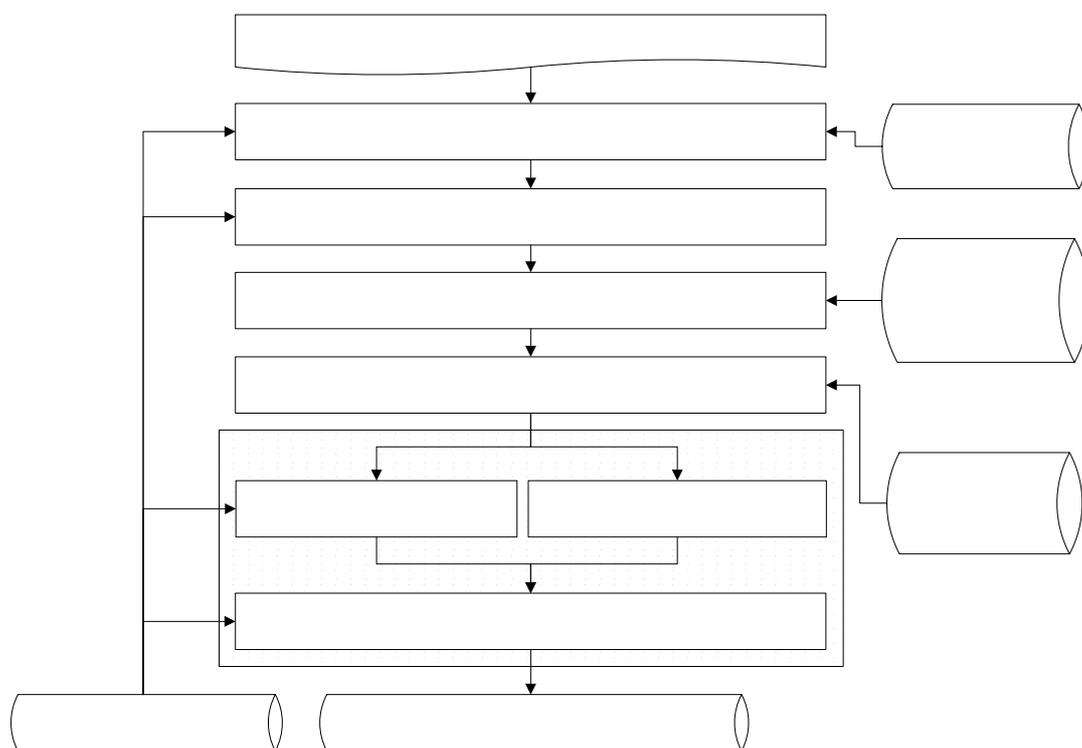


Figure 1. Flow chart for the CWAC system

Process 1. System and non-word dictionaries checking: If the n-gram input can be found in the system dictionary, it will be labeled as a **System-word**. If the n-gram input can be found in the non-word dictionary, it will be labeled as a **Non-word**. The non-word dictionary was created from the non-words of the generated n-grams of **ASBC** sentences, exclusive of the 10,000 testing sentences. The Top 10 most frequently n-grams in the non-word dictionary are: 一個, 這個, 就是, 這種, 他的, 上的, 爲了, 都是, 這是 and 也是.

Process 2. Segmentation by the system dictionary: In this stage, the n-gram input will be segmented by two strategies: forward (left-to-right) longest word first, and backward (right-to-left) longest word first. Note that the “longest syllabic word first strategy” is an effective technique for Chinese word segmentation. If forward and backward segmentations of the n-gram input are different, the CWAC system will be triggered to compute the products of all word lengths for these segmentations. If both products are equal, the backward segmentation will be selected. Otherwise, the segmentation with the greatest product will be selected. According to our previous experiments [Tsai *et al.* 2004], backward segmentation precision is, on average, 1% better than that of forward. As an example, the forward and backward segmentations of 將軍用的毛毯 are /將軍/用/的/毛毯/ and /將/軍用/的/毛毯/, respectively. Since both products are equal ($2 \times 1 \times 1 \times 2 = 1 \times 2 \times 1 \times 2$), the selected segmentation output for this process is /將/軍用/的/毛毯/ as it is the backward one. For clarity, the segmentation output from **Process 2** is referred to as *segmentation2*.

Process 3. Stop word checking: In this stage, all words in *segmentation2* will be compared with the stop word list. There are three types of stop words: **beginning**, **middle**, and **end**. The stop word list used in this study is given in Appendix A. These stop words were selected by human editors. If the first and last words of *segmentation2* can be found on the list of beginning and end stop words, they will be eliminated from the *segmentation2*. For those cases in which the word number of *segmentation2* is greater than two, middle stop word checking will be triggered. If a middle stop word in *segmentation2* can be found, the n-gram input will be split into new strings at any matched middle stop word. These new strings will be sent to **Process 1** as new n-gram input. For example, *segmentation2* of the n-gram input 但可怕的腸病毒啊 is /但/可怕/的/腸/病毒/啊/. Since there is a middle stop word “的,” a beginning stop word “但,” and an end stop word “啊” in this *segmentation2*, the new strings “可怕” and “腸病毒” will be sent to **Process 1** as new n-gram input.

Process 4. Non-word and Word checking based on linguistic knowledge: Once *segmentation2* has been processed by **Process 3**, the result is *segmentation3*. In this stage, the n-gram input will be auto-confirmed as **Non-word** only if *segmentation3* fits one of the following conditions:

- (a) it is a combination of a measure and a noun, such as 位/學生(measure word for student/student),
- (b) it is a combination of a number and a measure, such as 十/位(ten/measure word for ten),
- (c) it is a combination of a number, a measure and a noun, such as 十/位/

- 學生(ten/measure word for student/student),
- (d) it includes at least one non-Chinese character, such as @/公司 (company),
 - (e) its last word is a Chinese number (i.e., the *Neqa* and *Neu* words in the CKIP lexicon [CKIP1995]), such as 擁有/三(own/three), and
 - (f) it contains at least one function word, such as 問題/一直 where “一直” is an adverb. We define adverbs, conjunctions and propositions as function words.

On the other hand, the n-gram input will be auto-confirmed as **Word** only if *segmentation3* fits one of the following conditions:

- (g) it is a combination of a single-character and a suffix, such as 停車 (stop/car), and
- (h) it is a combination of a prefix and a single-character, such as 第一 (the/first). The used suffixes and prefixes were mainly selected from the *Lists of Final-Bound and Start-Bound Complement* in [CKIP 1995].

For *process 4*, we have created a supporting database which is a collection of 31,089 noun-measure word-pairs based on the GuoYuRiBaoLiangCiDian (國語日報量詞典) [MDNA & CKIP, 1997]. The GuoYuRiBaoLiangCiDian is a Dictionary of Measure Words (DM) and a Collection Dictionary of Noun and Measure Words (CDNM). If a *segmentation3* can not be auto-confirmed as a **Non-word** or a **Word** in this process, it will be sent to *Processes 5a* and *5b*.

Process 5a. Part-of-Speech (POS) pattern checking: If the word number of *segmentation3* is two, POS pattern checking will be triggered. The CWAC system will first generate all possible POS combinations of the two words using the system dictionary. If the number of generated POS combinations is one and that combination matches one of the POS patterns (N/V, V/N, N/N, V/V, Adj/N, Adj/V and Adj/Adj), the 2-word string will be sent to *Process 6*. This rule-based approach combines syntax knowledge and heuristic observation in order to identify compound words. For example, since the generated POS combination for *segmentation3* 食品/公司 is N/N, 食品公司 will be sent to *Process 6*.

Process 5b. Polysyllabic word checking: If both first and last words of a *segmentation3* are polysyllabic words and its character number is less than or equal to five, the *segmentation3* will be confirmed as a **Word**; otherwise, it will be sent to *Process 6*. Take segmentation /東港/黑/鮪魚/ as example. Since its first word “東港” and last word “鮪魚” are polysyllabic words and its character number is five, it will be auto-confirmed as a **Word** and will be sent to the on-line dictionary.

Process 6. Segmentation ambiguity checking: This stage consists of four steps.

- 1) Thirty randomly selected sentences that contain the n-gram input will be extracted from a large-scale corpus. The details of large-scale corpus used in this study will be addressed on subsection 3.1. As an example, the Chinese sentence 人人做環保 is a selected sentence for the n-gram input “人人”. If the number of the selected sentences is zero, this *segmentation3* will be confirmed as a **Non-word**.
- 2) These selected sentences will be segmented using the forward and backward longest word first techniques.
- 3) For each selected sentence, if the *segmentation3* can not be found in both forward and backward segmentations, the sentence will be treated as an *ambiguous sentence*. Take the n-gram input “用毛” and the selected Chinese sentences 將軍用毛毯 as example. This sentence will be treated as an ambiguous sentence for “用毛” since its *segmentation3* “/用/毛/” can not be both found in the forward segmentation /將軍/用/毛毯/ and backward segmentation /將/軍用/毛毯/.
- 4) If the ambiguous ratio of *segmentation3* is less than or equal to 50%, it will be confirmed as a **Word**; otherwise, it will be confirmed as a **Non-word** (As per our observation, the ambiguous ratios of words are usually less than or equal to 50%).

3. Experimental Results

The experiment is designed to show the UWI performance of our CWAC system on n-gram frequency ≥ 2 based on the *Segmentation Standard in Taiwan* [CKIP 1996].



Figure 2. An example of three matching sentences for the tri-gram “腸病毒” from the **OPENFIND** search results

3.1 Large-scale Corpus

By a large-scale corpus we mean one whose texts are collected from extensive Chinese web sites. We select the cached Chinese web pages of **OPENFIND** [OPENFIND], one of the most popular Chinese search engines, to be our large-scale corpus. For the example in Figure 2, CWAC extracts three matching sentences 腸病毒感染症, 談腸病毒(含克沙奇病毒)感染 and 腸病毒感染症簡介 for the tri-gram “腸病毒” from the cached **OPENFIND** search results.

3.2 The Experimental Results

The experiment is conducted as follows. First, we randomly select 10,000 sentences from **ASBC** 1.0 [CKIP 1995] to form a test sentence set. Then, we extract all n-grams with frequencies ≥ 2 from this test sentence set. The word precision of these extracted n-grams is 52%(=2623/5045). Since **ASBC** is a segmented corpus, we can extract a word set from the test sentence set. Finally, according to this word set and the CWAC auto-confirmed word set, we compute the UWI performance of our CWAC system as shown in Table 1. The computation of UWI performance excludes those n-grams labeled as **System word**. Table 1 shows that our CWAC system achieves (89.66%, 97.14%, 96.95) precisions, (66.67%, 86.72%, 86.12%) recalls for n-gram frequencies of = 2, ≥ 3 and ≥ 2 , respectively, for the selected sentences.

Table 1. The UWI performance of the CWAC system using the cached OPENFIND corpus for n-gram frequency of = 2, ≥ 3 and ≥ 2 based on the *Word Standard in Taiwan* [CKIP 1996].

n-gram Frequency	Precision	Recall	F-measure
= 2	89.66 (52/58)	66.67 (52/78)	76.47
≥ 3	97.14 (2207/2272)	86.72 (2207/2545)	91.63
≥ 2	96.95 (2259/2330)	86.12 (2259/2623)	91.21

4. Conclusion and Directions for Future Research

In this paper, we create a CWAC system that uses a LFSL (linguistic approach first, statistical approach last) hybrid approach to auto-confirm whether an n-gram input is a word. Our experiments show that this approach can effectively achieve a better than 97% precision and 86% recall for automatically confirming n-grams as words. In previous approaches, the hybrid UWI system of Chen [Chen *et al.* 2002] has a precision of 89%, and a recall of 68%. Sun [Sun *et al.* 2003] reported a precision of 84.28% and a recall of 83.53% for the name entity (NE) identification. Keep in mind that these approaches used different corpora and a fair comparison would be quite difficult. There are two kinds of errors that a CWAC system could make: 1.

Non-word error: when an n-gram is a word but auto-confirmed as a **Non-word** and 2. *Word error*: when an n-gram is not a word but auto-confirmed as a **Word**. According to experimental results, 96.2% *Non-word errors* and 88.8% *Word errors* occur in *process 6* of the CWAC system. Thus, a major focus for improving our CWAC system will be on the output of *process 6*.

This method is our first attempt to create a CWAC system. To improve the performance of *process 6*, some statistical word boundary checking (or machine learning) methods, such as local maxima algorithm [Silva *et al.* 1999], will be investigated in our future work. We have also considered a building-block approach to construct a multi-CWAC system. Creating an on-line Chinese word auto-identification (CWAI) system comprised of a Chinese word auto-detection (CWAD) system will be a project in the future.

References

- Chen, K.J. and W.Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceedings of 19th COLING 2002*, Taipei, pp.169-175
- Chieu, H.L. and H.T. Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," *Proceedings of 19th COLING 2002*, Taipei, pp.190-196
- Chang, J.S. and K.Y. Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction," *International Journal of Computational Linguistics & Chinese language Processing*, 1997
- Chen, K.J. and S.H. Liu, "Word Identification for mandarin Chinese Sentences," *Proceedings of 14th COLING*, 2002, pp. 101-107
- Church, K.W., "Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to $p/2$ than p^*p ," *Proceedings of 18th COLING 2000*, pp.180-186
- CKIP (Chinese Knowledge Information Processing Group), *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Taiwan, Taipei, Academia Sinica, 1995.
- CKIP (Chinese Knowledge Information Processing Group), *A study of Chinese Word Boundaries and Segmentation Standard for Information processing (in Chinese)*. Technical Report, Taiwan, Taipei, Academia Sinica, 1996.
- Dunnin, T., "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, no. 1, 1993
- Lai, Y.S. and C.H. Wu, "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-based Likelihood Ratio," *International Journal of Computer Processing Oriental Language*, 13(1), pp.83-95
- Ma, W.Y. and K.J. Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," *Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp.31-38
- MDNA (Mandarin Daily News Association 國語日報出版社) and CKIP (中央研究院詞庫小組), *GuoYuRiBaoLiangCiDian (國語日報量詞典)*, 1997 (in Chinese)
- MPCM (Mandarin Promotion Council Ministry of Education in Taiwan), *ZhongBi-anGuoYuCiDian (重編國語辭典)*, <http://140.111.1.22/clc/dict/>, 1998 (in Chinese)
- OPENFIND, OPENFIND Chinese Search Web Site, <http://www.openfind.com.tw/>

- Smadja, F., "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, 22(1)
- Silva, J.F., Lopes, G.P., "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units," *In Proc. of the 6th Meeting on the Mathematics of Language*, 1999, pp.369-381
- Sproat, R., C. Shih, W. Gale and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3), 1996, pp.377-404
- Sun, J., M. Zhou and J. Gao, "A Class-base Language Model Approach to Chinese Named Entity Identification," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 2, August 2003, pp.1-28
- Tsai, J. L., W. L. Hsu and J. W. Su, "Word sense disambiguation and sense-based NV event-frame identifier," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 7, No. 1, February 2002, pp.29-46
- Tsai, J. L., W. L. Hsu, "Applying NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem," *Proceedings of 19th COLING 2002*, Taipei, pp.1016-1022
- Tsai, J.L., Hsieh G and W.L. Hsu, "Auto-Generation of NVEF Knowledge in Chinese," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 1, February 2004, pp.1-24
- Zhang, H.P., Q. Liu, H.K. Yu, X.Q. Cheng and S.B. Zhang, "Chinese Named Entity Recognition Using Role Model," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 2, August 2003, pp.29-60

Appendix A. Stop Words List

I. Begining stop word list

/兒/呀/嗎/吧/呢/呼/了/是/你/我/他/又/等/既/或/有/到/去/在/爲//及/和/與/之/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/能/將/此/可/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但//別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚//皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/啲/等/著/嗎/嗨/嗚/噲/愈/跟/較/過/嘛/嘎/噃/嘻/嘿/噓/撲/罷/噹/噯/還/雖/嚕/

II. Middle stop word list

/可/已/各/被/到/等/既/但/且/而/並/同/又/爲/是/有/或/及/和/與/之/的/在/的/在/以/已/將/與/和/是/及/也/或/之/於/由/都/並/卻/且/只/則/但/又/才/仍/該/各/其/有/時/前/後/上/中/下/再/更/不/很/最/多/非/稍/否/至/了/吧/嗎/但/因/爲/而/且/就/對/雖/裡/裏/等/要/把/到/去/給/打/做/作/個/你/妳/我/他/她/它/們/這/那/此/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/能/將/此/可/與/到/向/以/用/乃/入/又/下/久/乎/者/已/互/仍/勿/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/但/別/即/吧/呀/更/沒/矣/並/呢/或/所/則/卻/哉/很/後/怎/既/甚//皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/啲/等/著/嗎/嗨/嗚/噲/愈/跟/較/過/嘛/嘎/噃/嘻/嘿/噓/撲/罷/噹/噯/還/雖/嚕/

III. End stop word list

/等/及/與/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/能/將/此/可/會/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但/別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚//皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/啲/等/著/嗎/嗨/嗚/噲/愈/跟/較/過/嘛/嘎/噃/嘻/嘿/噓/撲/罷/噹/噯/還/雖/嚕/
