Computer-Assisted Item Generation for Listening Cloze Tests and Dictation Practice in English

Shang-Ming Huang[†], Chao-Lin Liu[†], and Zhao-Ming Gao[‡]

[†] Dept. of Computer Science, National Chengchi University, Taipei, Taiwan
[‡] Dept. of Foreign Languages and Literatures, National Taiwan University, Taipei, Taiwan
chaolin@nccu.edu.tw, zmgao@ntu.edu.tw

Abstract. We take advantages of abundant text resource on the Internet and information about English phonetics for assisting human teachers to prepare test items for listening and dictation in English. In this preliminary exploration, we built an environment in which teachers choose words that they want to have test items for, and teachers compose the final test items based on the test items that are algorithmically generated by our system. The output of the current system indicates that computers can play active roles in assisting the composition of test items, though we have not done a field test over the usability issues.

1 Introduction

Traditional wisdom dictates that listening, speaking, reading, and writing are four major components in learning a language. Listening is a very important input channel for everyday communication, and so for learning languages. Listening is not just an important channel for learning languages, but also a skill that needs to be learned [1]. As a result, both teachers and learners need to have ways to evaluate the competence in listening. A dominant way of evaluating competence in listening is the listening comprehension tests [2], in which examinees answer questions that are related to a spoken segment of story or conversation. Tests of this form allow test administrators to evaluate students' high-level understanding of the spoken passage. In contrast, listening cloze tests (cf. [3]) examines students' capabilities in comprehending the sounds of particular words, and is a mechanism for evaluating students' low-level listening capabilities. Although few will deny the importance of gauging students' high-level and low-level competence in listening [4, 5], there is an imbalance in the availability of practice test items for listening comprehension and listening cloze. This is probably due to the fact that most real-world examinations employ listening comprehension tests rather than listening cloze tests, and the market has a much stronger demand on the practice material for listening comprehension.

Computer-assisted item generation [6] is not a brand new idea for the study of educational technologies. Although test items constructed by human experts are highly preferred for their quality, it is also costly to create such high-quality items while keeping the security of the item pools. When relying on a limited number of human experts to construct test items, the coverage of the constructed test items may be confined to the

R.W.H. Lau et al. (Eds.): ICWL 2005, LNCS 3583, pp. 197 – 208, 2005.

interests and views of these item creators, consequently impairing the quality of the test items. For computer assisted language learning, the Web serves as a good source of text files, and is already helping language learners around the world. Given the ample text files on the Web, we can extract sentences from the Web, compile the extracted sentences into candidate test items, and allow the teachers to select and edit the candidates before delivering the test items for students' practice and evaluation.

Not surprisingly, researchers have applied techniques of natural language processing in this computer-assisted language learning task. For instance, Stevens [7] applies the concept of concordance and Coniam [8] considers the concept of word frequency in generating vocabulary test items from corpora. Wang et al. [9] consider multiple linguistics-based techniques, including selectional preferences and word sense disambiguation [10], in assisting the generation of high quality multiple-choice cloze items. In addition to finding better ways for providing learning material, researchers have also explored the possibility of conducting higher level analyses of learners' language proficiency with computer techniques. Michaud et al. [11] propose methods for analyzing the grammatical structures of written English, and Burstein et al. [12] implement a system for evaluating essays of examinees of the TOEFL tests.

We propose computational methods that help teachers of English to prepare test items for listening cloze, and hope to offer learners of English a middle ground for diagnosing their own deficiency in listening capabilities. A listening cloze item is similar to an ordinary cloze item, except that the test item is delivered in the audio form. Hence, in addition to constructing a sentence for a test item, our system must also recommend distractors for the keywords that are deleted from the main sentences. In addition, we report exploration of classifying errors found in learners' dictation of English sentence. Dictation is relatively more difficult than multiple-choice test items, and implementing this type of testing environment grants ourselves the possibility of constructing a practice environment that may provide appropriate material that adapts to students' different levels of proficiency in English [13]. Similar to Coniam's Text Dictation system [14], we employ the concept of partial dictation [15], in which students are required to put down part of the spoken sentences. Unlike Coniam's rule-based method for giving partial credits to students' dictations, we classify students' answers in order to help students identify their weakness in the dictation task.

We provide an overview of our system in Section 2, elaborate on how we apply phonetics to generate listening cloze test items in Section 3, and present a way to evaluate student's dictation using some linguistics-related criteria in Section 4.

2 System Overview

At the current stage, we consider three different levels of learning English vocabulary. The most popular form of multiple cloze tests, illustrated below, should be the easiest form. In this type of cloze tests, examinees will choose the best candidate from the pro-vided alternatives for the deleted word in the given sentence. As the examinees can read all the available information in the test item, this form of tests is considered to be relatively easier than the listening cloze and dictation tests.

All the flights to and from Hong Kong were ___ because of the heavy thunderstorm.
(A) advised (B) disclosed (C) cancelled (D) benefited

In previous work [9], we build a system that takes advantages of resource available on the Web for assisting teachers to compose multiple-choice cloze items. Figure 1 shows the block diagram for this system. We download and preprocess text files from the Web, augment such information as part-of-speech and root forms of inflected words with the words, and save the results in the tagged corpus. Applying techniques for word sense disambiguation, our system allows teachers to request sentences that include a specific word with a specific meaning in the sentence, and choose appropriate distractors for the keyword to form a multiple-choice cloze item, while ensuring that there is only one correct answer to the composed item.

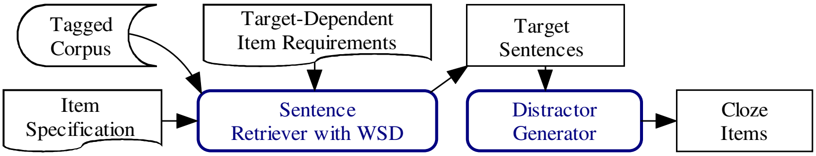


Fig. 1. Block diagram for generating multiple-choice cloze items [9]

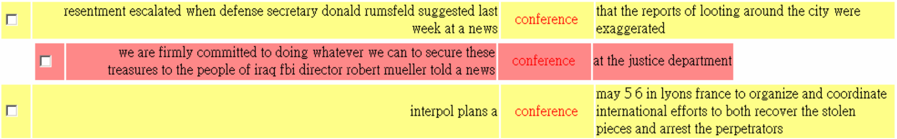
We use our Web crawler to obtain raw material from the Internet [16]. The text fragments are segmented into sentences and saved into the database. The corpus has 163,719 sentences, which include a total of 3,077,474 words that comprise 31,732 different words. Recognizing the current choice of the text sources may not be the best possible choice, we use them simply because they offer satisfactory quality of articles and free accessibility. Our main goal is to demonstrate the viability of our system architecture. Whenever necessary, we can replace the text corpora with better ones.

In this work, we continue to use the tagged corpus as the source of sentences for ex-tending our system to assist the tasks of dictation and multiple-choice listening cloze tests. In Section 3, we focus on how we choose distractors for the listening cloze tests. In Section 4, we turn our attention to how we may evaluate students' competence in dictation.

We look forward to a vertically integrated system that is capable of adaptively interacting with students for assisting the evaluation of competence in English vocabulary, putting the resulting components together. The previous work on cloze tests allows learners to examine their English vocabulary in reading form, and the results re-ported in this paper extend the scope to listening and active production in dictation.

3 Generating Multiple-Choice Listening Cloze Items

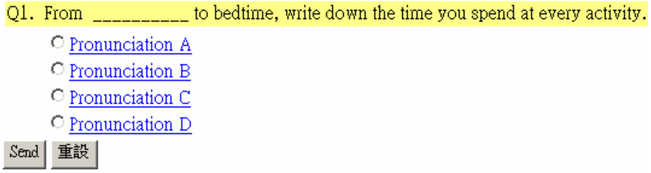
Our system helps teachers create multiple-choice listening cloze test items. First, teachers request sentences for a selected keyword, and our system will search the corpus for sentences that include the keyword. Our system will arrange the sentences to facilitate the selection of sentences by aligning the keywords in the same column.



□	resentment escalated when defense secretary donald rumsfeld suggested last week at a news	conference	that the reports of looting around the city were exaggerated
□	we are firmly committed to doing whatever we can to secure these treasures to the people of iraq fbi director robert mueller told a news	conference	at the justice department
□	interpol plans a	conference	may 5 6 in lyons france to organize and coordinate international efforts to both recover the stolen pieces and arrest the perpetrators

Fig. 2. Teachers select appropriate candidate sentences from the list

This is the same as a concordancer, which is shown in Figure 2. In addition, our system attempts to put sentences that have "similar phonetic environment" around the keyword, which will be explained in Section 3.1. Teachers then select appropriate sentences from these arranged sentences, and our system will automatically select the distractors from our databases and compile the multiple-choice problems.



Q1. From __________ to bedtime, write down the time you spend at every activity.

○ Pronunciation A
○ Pronunciation B
○ Pronunciation C
○ Pronunciation D

Fig. 3. A format of the multiple-choice listening cloze item

There are two different types of multiple-choice problems in our system. The first type tests on individual keywords, while the other type removes the keyword and the words that immediately precedes and follows the keyword. Figure 3 shows one format of the generated multiple-choice listening cloze items. The keyword in the test item is replaced with underscores, and there are four options for the test item. Clicking on the hyperlinks, students will hear the audio for each distractor. At this moment, we create the audio for the words and sentences with the AT&T Natural Voice [17], and can replace the audio files with human-recorded files for test items of higher quality. There is only one correct option among the alternatives, and the other distractors are automatically generated by our system. Students can click on one of the four checkboxes. and submit their answers.

A more difficult format of the listening cloze tests, which is not shown here, does not show the whole sentences immediately. Instead, students check a button to listen to the whole sentence for a test item up to three times, but they do not have to listen to the whole sentence three times before answering the question. Students will not see the printed form of the test item, which are in the same format as the one shown in Figure 3, until entering the answering mode. After entering the answer mode, the audio for the whole sentence will not be repeated. This format is more difficult because students have to listen to the whole sentence before they answer. In contrast, the test item shown in Figure 3 allows students to focus solely on the sounds of the keywords.

In addition to controlling the item difficulty by manipulating how and when students can listen to the recordings, modulating the playing speeds of the recordings is another conceivable alternative. With appropriately supported software techniques, we can change the speeds of both human-articulated and machinesynthesized recordings. Our current choice of using the AT&T voice synthesizer is a result of an arbitrary decision, though the synthesizer does produce satisfactory recordings for short text segments.

For facilitating the selection of distractors, we employ two metrics for measuring the similarity between words. We consider the features of phonemes that are contained in words for computing the distances between words, and construct a database based on our definition of the *phonetic distance* (**PD**). Words that are similar in terms of the phonetic distance are clustered in the PD database. We also construct a database based on our definition of the consonant-vowel patterns (**CVP**s). Words that have the same consonant-vowel patterns are clustered for future use. We provide more details about the PD and CVP databases in Sections 3.2 and 3.3, respectively.

When composing a test item, our system randomly chooses from the PD database those words that are similar to the keyword first. If there are less than three words available in the PD database, more words will be retrieved from the CVP database. If more distractors are needed, our system will create words to make up the distractors. Disregarding whether the words do exist, we randomly replace vowels of the keyword with other vowels to make a new word. Consider the word “structure” which has two vowels ‘u’. We may select the first vowel ‘u’ and randomly replace it with other vowels like ‘a’, ‘e’, ‘i’ or ‘o’, and come up with candidate distractors like “stracture”, “strecture”, “stricture” and “stracture”, respectively. Although these distractors may not be appropriate in a reading cloze test, the synthesized sounds of these created words may be quite distracting to students who are not good at listening.

3.1 Phonetic Concordance

The actual sound of words may be influenced by its context, so we try to put candidate items whose keywords have similar phonetic contexts near each other. For example, the sentence “You can get them at the drugstore”. The sound of the word “get” is influenced by the phoneme /□/ of the word “them” so that the phoneme /t/ may not be pronounced clearly, and students may get confused by these variations. Considering this situation, we believe that it is useful to arrange candidate items that have similar phonetic environments together to better serve the item preparation process.

To this end, we consider the syllables that are immediately before and after the keyword in determining whether a keyword has similar phonetic contexts in two sentences. For each candidate item, we obtain these syllables from the words that immediately precede and follow the keyword, and concatenate the syllables with the phoneme sequence of the keyword. We compute the similarity between two phoneme sequences with a minimum edit distance (MED) algorithm [cf. 18], which employs the definitions of phonetic distance that we explain in Section 3.2. We consider sequences similar if their phonetic distance is no larger than 50.

For example, consider two sentences: “The sergeant simply abandoned his position”, and “This myth was effectively abandoned in the 1990s”. The phonemic environments of the word “abandoned” in these sentences are: “-ly abandoned his” and “-ly abandoned in”. The corresponding phoneme sequences are: “-ly-a-ban-don-ed-his” and “-ly-a-ban-don-ed-in”. After computing the phonetic distances between the sequences, our system will find that the phonemic environments of “abandoned” are similar in these sentences, and will put them together in the interface that looks like Figure 2. In Figure 2, the phonetic environment of the word *conference* in the indented sentence (in red background) is similar to that in the first sentence.

3.2 Phonetic Distance

Table 1. Phonetic features of three phonemes

Phoneme	Place of Articulation	Manner of Articulation
/m/	bilabial	nasal
/n/	alveolar	nasal
/d/	alveolar	stop

Each phoneme has its phonetic features [19, 20]. The number of common features determines the similarity between phonemes.We have selected two features which are commonly used to differentiate phonemes: the place of articulation and the manner of articulation, although it requires more than two features to completely tell all the phonemes apart. Table 1 shows selected phonetic features of the phonemes /m/, /n/, and /d/: /m/ and /n/ share a common feature, and /m/ and /d/ differ in both features. Perceptually, /m/ is more similar to /n/ than to /d/. This example suggests that the number of different phonetic features reflects our perception about different phonemes.

Hence, we define the *phonetic distance*, PD, between two phonemes based upon the number of the different phonetic features between them as follows. According to the definition, the phonetic distance between /m/ and /n/ is 2, while the phonetic distance between /m/ and /d/ is 3. Distances are defined as 1 for different phonemes which share the same place and manner of articulation.

- PD is 0 if the consonants are exactly the same.
- PD is 1 if both the place and the manner of articulation are the same.
- PD is 2 if either the manner or the place of the articulation is different, but not both.
- PD is 3 for all other cases.

Similarly, we select three features for defining the distances between vowels: the position of the tongue, vowel height (the highest part of the tongue), and the shape of the lips. The definition for phonetic distance between two vowels follows.

- PD is 0 if the vowels are exactly the same.
- PD is 5 if the three features are exactly the same.
- PD is 10 if one of the three parameters is different.
- PD is 15 if two of the parameters are different.
- PD is 20 if all of the three parameters are different.

For example, the phonetic features of the vowel /i/ are: the tongue is high, the tongue position is toward the front, and the mouth is unrounded [20]. The phonetic features of /□/ are: the tongue is semi-high, the tongue position is toward the front, and the mouth is unrounded. There is only one different feature, so the phonetic distance between /i/ and /□/ is 10. For another vowel /u/, the phonetic features are: the tongue is high, the tongue position is toward the back, and the mouth is rounded. There are two different features between /i/ and /u/, so their phonetic distance is 15.

Perceptually, the difference between vowels is much more obvious than that between consonants. Most of the time, vowels are also pronounced louder and more prominent than consonants. To reflect these facts, we amplify the distances between vowels. As consonants are very different from vowels, we define the phonetic distance between any vowel and any consonant as 50.

We define the distance between two words as the minimum edit distance [cf. 18] that is calculated with the phoneme sequences of the words. We first convert the words into their phonemes by looking into the Merriam-Webster online dictionary [21] and grapheme-phoneme-conversion rules, and split the phonemes into individual symbols with standard techniques of lexical analyzers in Computer Science [22]. We then apply the standard MED algorithm for computing the phonetic distance between the words. We apply the aforementioned phonetic distances when encountering insertion, deletion, and substitution of phonemes in the calculation. (It is understood that the actual sound of an individual phoneme may change because of its contextual phonemes [19]. The fact that we use the dynamic algorithm-based MED algorithm is tentative to ignoring such contextual influences.)

Figure 4 shows the process of computing the phonetic distance between *intercately* and *intricately* with the MED algorithm. The leftmost column and the bottom row in the *edit-distance matrix* show the phoneme sequences of he words being compared. In Figure 4, phonemes of *intercately* and *intricately* are on the leftmost column and the bottom row, respectively. Each cell in the matrix reflects the minimum phonetic distance between the corresponding phoneme sequences which appear on the leftmost column and the bottom row, e.g., PD(intricately, intercately)=17 and PD(intricately, inter)=122. The underlined bold cells show the shortest distances between substrings of *intercately* and *intricately*.

ɪ	285	159	154	112	62	78	43	69	67	17
l	285	109	104	62	103	28	75	19	17	67
t	235	107	102	60	53	25	67	17	19	69
ə	185	105	102	58	23	67	17	67	111	79
k	170	55	52	8	55	17	67	61	64	114
ɚ	120	52	50	5	17	67	59	109	154	122
t	100	2	0	2	52	54	104	104	107	157
n	50	0	2	4	54	57	107	109	111	161
ɪ	0	50	100	150	150	200	215	265	315	315
	ɪ	n	t	r	ɪ	k	ə	t	l	ɪ

Fig. 4. PD between intricately and intercately

In an earlier attempt, we arbitrarily set 50 in PD as the boundary for similar words. When there are two extremely different vowels and three extremely different consonants between the words being compared, the phonetic distance will be 2*20+3*3=49. Hence, a phonetic distance that is larger than 50 indicates really different words. This absolute threshold may not work very well for short words, though.

In the most recent experiments, we have found that a good choice of threshold should depend on the number of syllables in the words being compared. Using a large threshold will admit very different phones to be treated as similar phones, and makes the resulting test item less challenging. Using a very small threshold might cause two kinds of problems. There could be insufficient similar words for each entry in the PD database for even one single test item. Worst yet, when the previous situation does occur, these very few candidates will be repeatedly used, making the resulting test items less useful in practice. Making the threshold a function of the number of syllables is more flexible, but has not been implemented in the most recent version of our system.

Using the phonetic distance, we construct the PD database. We compute and save a list of similar words for each word, except for the function words (such as articles, prepositions, conjunctions, auxiliary verbs and pronouns, etc.). Words that are less than 50 apart in terms of the phonetic distance are considered similar. The wordlist for each word in the database will be used in the selection of distractors for the multiple-choice items.

3.3 Consonant-Vowel Patterns

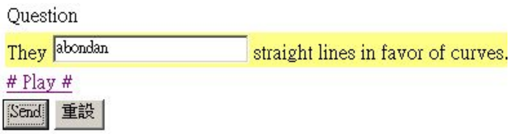
In addition to the phonetic distance, we also compare the consonant-vowel patterns (CVPs) of two words to determine their similarity. After converting words into phoneme sequences by consulting the on-line dictionary, we convert a vowel into the symbol “+” and a consonant into the symbol “-”. Hence the consonant-vowel pattern of the word “follow” is “-+-+-”, and the pattern of “hollow” is also “-+-+-”.

Words that have the same CVP may have similar pronunciations, e.g., “follow” and “hollow”, and can be used as distractors for each other. However, there are words that have the same CVP, but they have a large phonetic distance, e.g., “absolute” and “organic”. Some words have different CVPs, but their phonetic distance is close, e.g., “absolute” and “calculate”. Therefore, the information about the consonant-vowel patterns does not guarantee the similarity between words.

Nevertheless, for having a fallback for the PD database, we cluster words that have the same consonant-vowel pattern into groups in the CVP database. We convert each word in the corpus into their consonant-vowel patterns. Words with the same consonant-vowel patterns are clustered, and the information about the consonant-vowel clusters is used in generating the multiple-choice items when necessary.

4 Dictation Error Analysis

Teachers create items for dictation with a procedure that is similar to that for creating multiple-choice listening cloze items. Figure 5 shows a format of the created dictation test item. The sentence with a blank text field is the test item, where students are supposed to fill out the missing word. Clicking on the hyperlink “#Play#” below the test item, students will hear the whole sentence. Similar to how we create audio for the listening cloze items, we rely on the AT&T Natural Voice for synthesizing the audio for the dictation items. Also similar to the listening cloze items, there is a corresponding, more difficult format of the test, in which students will have to listen to the whole sentence before they can see the sentence with the deleted word as is shown in Figure 5. After filling out the text, students submit their answers by clicking on the “Send” button. Our system will save the answers and deliver the next item.



Question

They abondan straight lines in favor of curves.

Fig. 5. An item for dictation

As an attempt to formulate a principled model of the errors committed by real-world students in dictation, we collected and analyzed actual students’ dictations from the English classes of the last author. Unfortunately, we must admit that it is rather hard to come up with a list of well-founded explanations for the observed errors. Sometimes, students put down the correct words in wrong tenses or inflected forms. Sometimes, students seemed to know what they heard but could not spell the words correctly.

Although we cannot establish psycholinguistic foundations for the committed errors, we try to classify the errors purely from the linguistic viewpoints. Specifically, we observed three types of errors, and refer these types as *syntactic errors*, *spelling errors*, and *phonetic errors* for convenience. Section 4.1 explains how we classify the errors, and Sections 4.2 provides details about how our system provides feedback to the students based on their types of errors.

4.1 Three Types of Errors

Students' answers that included **morphological errors** are actually very close to the correct answers. For instance, students may have put down "maps" and "combine", respectively, for "map" and "combined." Errors of this type can be detected with a lemmatization technique [cf. 10]. Lemmatization is the process of converting inflected words to their original forms. For instance, results of lemmatizing "combined" and "maps" are "combine" and "map", respectively. We employ the Porter's Stemmer [23] to lemmatize students' answers and the correct answers, and compare the resulting strings. **Morphological errors** are indicated if some words in these strings become matched only after the lemmatization step, while these words do not match before being lemmatized.

We call the second type of errors **spelling errors**. Students may have heard the phones of the words, but cannot reconstruct the spellings of the pronounced words. It is common that, in this case, the dictation is similar to the correct spelling, but contains a few missing and/or redundant characters. We measure the similarity between such misspelled and correct words from two related perspectives: Maximum Common Characters (MCC) and Minimum Edit Distance (MED). Here we use the Levenshtein cost function for the MED algorithm [cf. 18], i.e., 1 for both insertions and deletions, and 2 for substitutions.

Table 2 shows the MCC and MED for some samples. The leftmost two columns show the correct answers and students' dictations. The third and the fourth columns show the MCC and MED values between the word pairs. Take "intricately" and "intercately" for example. Their MCC value is 10 because of the 10 common characters 'i', 'n', 't', 'r', 'c', 'a', 't', 'e', 'l', and 'y'. The MED value is 2 because of 1 deletion ('e') and 1 insertion ('i').

Table 2. The MCC and MED of some word pairs

Answer	Dictation	MCC	MED
intricately	intercately	10	2
interwoven	interwoved	9	2
with	intercately	2	11
civilizations	interwoved	2	10
inter	retni	5	8
inter	in	2	3

Let *L* be the number of characters in the correct answer, *A* and *D* be the correct answer and the dictation, respectively. After observing the collected data, we consider *D* is a misspelled form of *A* if **MCC(*A*, *D*) > *L*/2 and MED(*A*, *D*) < *L***. The first two instances in Table 2 show examples that meet the conditions, but the other instances do not.

Words that are considered to be a misspelled form of the correct words will be further analyzed for possible **phonetic errors**. As mentioned in Section 3.2, we use the MED algorithm to compute the phonetic distance between two words. By tracing back this optimal path from the upper right corner, we can identify the different phonemes between the words. Consider the edit-distance matrix of PD(intercately, intricately) in Figure 4. The optimal path is “17-17-17-17-17-17-2-0-0-0”. Drops in the numbers in the path reveal the different phonemes between the word pair. By this way, the different phonemes between two similar words can be found.

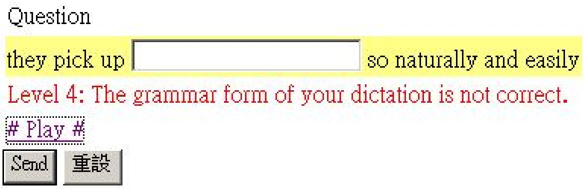
Recall, however, that the words that we compare in Section 3.2 are all correct words in English. Therefore, we can look up the online dictionary for their phones and produce corresponding phoneme sequences. When processing students’ dictations, it is not uncommon that students would put down *non*-existent words. It is not directly obvious how one can find the phones of these wrong words. To do as much as we can, we rely on word formation rules [24] and grapheme-to-morpheme rules [25] for converting students’ dictations into phoneme sequences. Although the correspondence between graphemes and phonemes are not perfectly regular [19, page 562], existing rules are not completely useless for converting non-existent words into phoneme sequences.

4.2 Item Feedback

A satisfactory feedback system should first identify the actual weakness in students’ competence, and provides the material that will really remedy the problems. Given that it is not easy to classify students’ errors in dictation, our exploration into providing feedback items for students is nothing but preliminary, if not bold. After classifying students’ errors based on the aforementioned three criteria, it is natural that our system responds to students’ dictations according to the classified error types.

- Level 1: the dictation is extremely different from the correct answer in spelling, suggesting that the student has no idea about the testing material
- Level 2: (spelling errors): the dictation is a misspelled form of the correct answer and the phonetic distance is larger than 50, suggesting that the student may have some idea about the testing material
- Level 3: (phonetic errors): the dictation is a misspelled form of the correct answer and the phonetic distance is no larger than 50, suggesting that the student has roughly caught the testing material
- Level 4: (morphological errors): the dictation is an inflected form of the correct answer, suggesting that the student may have exactly known the testing material
- Level 5: the dictation is exactly the same as the correct answer

In the prototype for assisting students to practice dictation, after classifying students’ answers, our system records the transactions in the students’ profiles, and continues to interact with the student. Figure 6 shows such a correspondence. In this example, our system determines that the dictation is not correct, and feedback this evaluation to the student. When necessary or requested, for level 3 errors, our system can show the phonemes that are correctly dictated. If students want to do more practice, the system will deliver more test items that are appropriate for the students’ competence levels.



Question

they pick up ______ so naturally and easily

Level 4: The grammar form of your dictation is not correct.

Fig. 6. Our system responds to a level 4 dictation

If the entered word is not perfectly correct, our system will continue the previous test item, and may repeat the same item for up to three times. After receiving three incorrect answers, our system will simply show the correct answer. If the entered is correct, i.e., level 5, our system allows the student to choose different types of new test items. Students can choose more items which test on the same keyword, items which test on keywords that have the same consonant-vowel pattern with the tested keyword, items which test on other keywords that have small phonetic distance with the tested keyword, or items which test on other items which have similar phonemic environments with the tested item.

5 Conclusions

We proposed a computer-assisted item generation system for helping teachers to create practice items for learning English vocabulary, and report the design of our system for generating listening cloze and dictation practice in this paper. We believe such functions will facilitate the construction of a Web-based test system for a large population. The current system offers aids in constructing test items in different levels of difficulty, including reading cloze, listening cloze to dictation, so our system is posed to support adaptive interaction with the students. Although the current efforts are preliminary, we believe that this is an important step toward realizing adaptive interaction with students in computer-assisted language learning.

We have identified some future work. As pointed out by reviewers, the system should be evaluated by teachers. Technically, we should consider contextual influence on the pronunciation of individual words, so MED is an imperfect method for computing difference between two words. Using a voice synthesizer, it should be possible to add more intermediate steps between listening cloze to dictation tests.

Acknowledgements

The authors thank Professor Maw-Kae Hor, Professor Berlin Chen, and anonymous reviewers for comments on improving a previous version of this manuscript. This work was supported in part by Grants 92-2213-E-004-004, 92-2411-H-002-061, 93-2213-E-004-004, and 93-2411-H-002-013 of the National Science Council, Taiwan.

References

1. Vandergrift, L.: Listen to learn or learn to listen, Annual Review of Applied Linguistics **24** (2004) 3–25
2. Mendelsohn, D.: Teaching Listening. Annual Review of Applied Linguistics **18** (1998) 81–101

3. UBC English Language Institute: http://www.eli.ubc.ca/teachers/lessons/speaking/presentation_skills/expository.html#c
4. Coniam, D.: Computerized dictation for assessing listening proficiency. Computer Assisted Language Instruction Consortium Journal **13**(2-3) (1996) 73–85
5. Ross, S.: Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. Language Testing **15**(1) (1998) 1–20
6. Irvine, S. H., Kyllonen, P. C. (eds.): Item Generation for Test Development, Lawrence Erlbaum Associates (2002)
7. Stevens, V.: Classroom concordancing: Vocabulary materials derived from relevant authentic text. English for Specific Purposes **10**(1) (1991) 35–46
8. Coniam, D.: A Preliminary inquiry into using corpus word frequency data in the automatic generation of English cloze tests. Computer Assisted Language Instruction Consortium Journal **16**(2-4) (1997) 15–33
9. Wang, C.-H., Liu, C.-L., Gao, Z.-M.: Using lexical constraints for corpus-based generation of multiple-choice cloze items. Proc. of the Seventh IASTED Int. Conf. on Computers and Advanced Technology in Education (2004) 351–356
10. Manning, C. D., Schütze, H.: Foundations of Statistical Natural Language Processing, MIT Press (1999)
11. Michaud, L. N., McCoy, K. F., Stark, L. A.: Modeling the acquisition of English: An intelligent CALL approach, Proc. of the Eighth International Conf. on User Modeling (Lecture Notes in Computer Science 2109) (2001) 14–23.
12. Burstein, J., Chodorow, M., Leacock, C.: Criterion^{SM}: Online essay evaluation: An application for automated evaluation of student essays. Proc. of the Fifteenth Annual Conf. on Innovative Applications of Artificial Intelligence (2003) 3–10
13. Brusilovsky, P.: Adaptive and Intelligent Technologies for Web-based Education. Künstliche Intelligenz, **13**(4) (1999) 19–25
14. Coniam, D.: Interactive evaluation of listening comprehension: How the context may help. Computer Assisted Language Learning **11**(1) (1998) 35–53.
15. Oller, J.: Language Tests at School: A Pragmatic Approach, Longman, London (1979)
16. China Post, http://www.chinapost.com.tw/; Studio Classroom, http://www.studioclassroom.com/; Taiwan Journal, http://taiwanjournal.nat.gov.tw/; Taiwan Review, http://publish.gio.gov.tw/fcr/
17. AT&T Natural Voice, http://www.naturalvoices.att.com/
18. Jurafsky, D., Martin, J. H.: Speech and Language Processing, Prentice Hall (2000)
19. Fromkin, V., Rodman, R., Hyams, N.: An Introduction to Language, Thomson Learning (2002)
20. International Phonetic Association: Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge Univ. Press, (1999)
21. Merriam-Webster OnLine: http://www.m-w.com/dictionary.htm
22. Levine, J. R., Mason, T., Brown, D.: Lex & Yacc, O'Reilly (1992)
23. Porter, M. F: An algorithm for suffix stripping. Program **14**(3) (1980) 130–137
24. Sinclair J. (ed.): Collins CoBuild English Guides: Word Formation, HarperCollins (1990)
25. Divay, M., Vitale, A. J.: Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. Computational Linguistics **23**(4) (1997) 495–523