Generating Confirmation to Distinguish Phonologically Confusing Word Pairs in Spoken Dialogue Systems

Kazunori Komatani **Ryoji Hamabe** **Tetsuya Ogata** **Hiroshi G. Okuno**

Graduate School of Informatics
Kyoto University
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan
komatani@i.kyoto-u.ac.jp

Abstract

The intelligibility of responses in spoken dialogue systems is important for communicating successfully, especially when the systems are used via cellular phones in noisy environments. We contrive language expressions of system responses to avoid users misunderstandings, which may occur when word pairs in a system's vocabulary phonologically resemble each other. We designed and developed a method to automatically add clarifying expressions when needed. Our method can be applied to various domains because no hand-written rules are needed in new domains. It exploits multiple knowledge sources such as dictionaries to get distinctive words as candidates to use in clarifying expressions. To select the best one, we introduce a criterion: ease in hearing. It is defined as the weighted sum of the uniqueness of the distinctive word and the difference between the distinctive and original words. We evaluated our method by applying it to the vocabularies of two systems. An experiment with five subjects showed that our system improved the intelligibility of phonologically similar words.

1 Introduction

Because of recent improvements in spoken dialogue systems, we can access information systems verbally with cellular phones [Komatani *et al.*, 2003; Raux and Eskenazi, 2004]. Since no special new apparatuses are needed when the telephone is used as the interface, such information services are expected to be used by various people, including the elderly.

Communication using speech inherently involves recognition errors. Phones, particularly cellular phones, often pick up background noise, so speech communication must be robust in such environments. Therefore, various studies such as [Hirsch and Pearce, 2000] have focused on improving speech recognition accuracy in noisy situations.

In addition to speech recognition errors, which correspond to the system mishearing the user, the possibility of the user mishearing speech should also be considered. Although the quality of text-to-speech (TTS) systems has improved, the control of intonation remains a problem. Automatically generating appropriate intonations and accents for individual proper nouns is nearly impossible. Therefore, the expressions a system uses should be intelligible so that the user easily understands them, especially when TTS systems are used.

We developed a method of automatically adding appropriate expressions to system responses that makes hearing content words including proper nouns easier. We aimed to reduce users mishearing and to improve the certainty of speech communications, even in noisy situations.

To clarify a confusing spelling, English speakers often use the phonetic code, e.g., "A for alpha", "B for boy", and "C for Charlie". Because we think that the phonetic code is not intuitive enough, we also used words or expressions that explain larger units than phones to indicate differences between confusing words. We used multiple knowledge sources (dictionaries) to automatically obtain candidate words to be added to confirmations as clarifying expressions.

Criteria are needed to select the best clarifying expression among candidates. We define the ease in hearing the clarifying expression using two criteria: the uniqueness of the clarifying expression and its difference from the original expression. By comparing weighted sums of the two criteria for each candidate, the best clarifying expression is selected and spoken to users. The candidates are automatically extracted from knowledge sources that are generally used, and the best candidate is automatically selected. This means that our method does not depend on a specific domain, so it can be applied to other domains without new generation rules.

2 Design of System Responses to Avoid User Misunderstanding

2.1 Misunderstanding Caused by Phonologically Similar Words

In communications using speech media, speech recognition errors cannot be avoided. To successfully communicate, user mishearing of the system's responses must be avoided, as must system recognition errors. In particular, errors in communicating content words are fatal in spoken dialogue systems because such words are used as keywords for queries. To avoid such errors, many systems require confirmation of content words [Hazen *et al.*, 2000]. The confirmation, however, may not be understood when phonologically confusing

S1: Please tell me your current bus stop, destination, or specific bus route.
U1: From *Kinkaku-ji* (金閣寺) temple.
S2: Did you say "from *Kinkaku-ji* (金閣寺) temple"?
U2: ??? (The user cannot discern whether the system said *Kinkaku-ji* or *Ginkaku-ji*.)

Figure 1: Example dialogue from conventional system.

S1: Please tell me your current bus stop, destination, or specific bus route.
U1: From *Kinkaku-ji* (金閣寺) temple.
S2: I am not sure whether you said *Kinkaku-ji* (金閣寺) temple or *Ginkaku-ji* (銀閣寺) temple. Does what you said begin with the character "*kin* (金)", which means gold in English?
U2: Yes.

Figure 2: Example dialogue from our system.

words are used. In such cases, humans can distinguish such words by focusing on the confusing part and adjusting the intonation, accent, speed, etc. For the systems, automatically giving proper nouns the proper intonation and accent is nearly impossible, just as non-native speakers have difficulty pronouncing unknown proper nouns correctly. Therefore, intelligible language expressions of the confirmations are needed as well as precise TTS engines to correctly transmit the system responses.

Our goal was to prevent the mishearing of content words in spoken dialogue systems. When the content words in confirmation responses are phonologically confusing, our system automatically adds expressions to clarify the words. The system automatically extracts phonologically similar word pairs from the vocabulary and generates clarifying expressions by using multiple knowledge sources. Clarifying expressions make confirmations intelligible, but may also make dialogues lengthy. What words will be added to the clarifying expressions can be controlled by setting a threshold to the phonetic distance between confusing words.

An example dialogue from a conventional system is shown in Figure 1, and one from our system is shown in Figure 2. Because phonologically confusing words are in the systems' vocabularies such as *Kinkaku-ji* and *Ginkaku-ji*, which are names of temples, the confirmation "Did you say from *Kinkaku-ji* temple?" in Figure 1 may sound similar to "from *Ginkaku-ji* temple", especially when the quality of the TTS is low or is listened to in a noisy environment. Therefore, our system generates a confirmation that contains an expression that clarifies the difference between phonologically confusing words (e.g., "Does what you said begin with the character *kin* (金), which means gold in English?" in Figure 2).

2.2 Generating Confirmation without Dependence on Domains

Our goal is to generate the kind of confirmation described in the previous subsection without dependence on any specific domain. That is, we do not manually describe rules for individual words in a system's vocabulary, but exploit existing knowledge sources such as dictionaries. This enables us to automatically apply our method to various systems.

Specifically, our system extracts word pairs that are phonologically similar from a system's vocabulary. It also automatically obtains candidates for clarifying expressions from dictionaries to distinguish the pairs. As a unit to explain the difference between confusing pairs, we adopted Chinese characters[1] because we think a larger unit than a phone is more intuitive. We exploit existing dictionaries that explain the unit, such as a set of words consisting of several Chinese characters, a Japanese-English dictionary, a dictionary describing readings of Chinese characters, and the Japanese phonetic code[2]. None of these dictionaries were constructed by hand for our system but generally used. The system automatically selects the best candidate expression based on its intelligibility. Thus, the clarifying expressions are generated automatically.

2.3 Intelligibility of System Responses

We assume that the intelligibility of system responses consists of two factors: difficulty and ease in hearing.

The difficulty indicates whether users understand the meaning instantly upon hearing an expression. Therefore, the intelligibility of a system response is improved when an expression with low difficulty is used. We define the difficulty of an expression by the difficulty of its words. We incorporated the difficulty by using dictionaries in which difficulty rankings are given.

The ease in hearing indicates whether the expression is confusing. Therefore, it is defined by considering both whether the expression is generally similar to other expressions and how the response changed phonologically from the original confusing response by adding the clarifying expression. We define the former as the uniqueness and the latter as the difference of a clarifying expression. The ease in hearing is defined by the weighted sum of the two criteria.

3 Generating Clarifying Expressions for Phonologically Similar Words

We describe our method for generating clarifying expressions by obtaining distinctive words from dictionaries. Confusing word pairs are automatically extracted from the system vocabulary and are distinguished using the distinctive words. The procedure is listed below and depicted in Figure 3.

1. We define the phonetic distance between two words. By calculating the distance for all the combinations in

[1] Because Chinese characters are ideograms, Japanese people often tell the difference between two words using the meanings of the characters.

[2] The Japanese phonetic code describes each Japanese syllable.

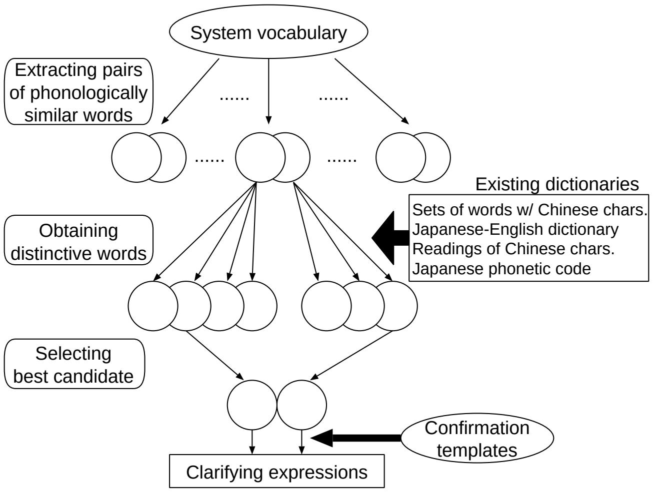

Figure 3: Flow to generate clarifying expressions.

the system's vocabulary[3], phonologically similar word pairs are extracted.

2. Distinctive words that include different spelling or phones between the phonologically similar pair are obtained for each pair from dictionaries. The words are candidates for use in clarifying expressions.

3. The best candidate among the distinctive words is selected. An expression of confirmation is generated by applying the candidate to templates for response sentences.

We describe the details of each step in the following sections.

3.1 Extracting Pairs of Phonologically Similar Words from System Vocabulary

We describe our method of extracting phonologically similar word pairs from a target system's vocabulary. First, we define the phonetic distance between words, which represents how phonologically similar two words are. Two words are phonologically similar, that is, confusing, if they have more common phonemes and less different phonemes. Furthermore, in Japanese, because words are pronounced by following a definite rhythm called mora [4], two words having the same number of morae tend to be confusing.

Considering these characteristics, we define a phonetic distance, $p.d.(w, w')$, between w and w' as follows:

$$p.d.(w, w') = \frac{e.d.(w, w')}{\log(\min(|w|, |w'|) + 1)}, \qquad (1)$$

where $|w|$ denotes the number of phonemes in word w. The numerator, $e.d.(w, w')$, denotes the edit distance[5] between the phoneme sequences w and w'. The edit distance is calculated using DP matching [Navarro, 2001]. In calculating an edit distance, a cost is introduced for each insertion, deletion, and substitution to reflect the mora structure. Here, C_1 is a set of vowels, choked sounds, and syllabic nasals, and C_2 is a set of consonants and semi-vowels. If edited phonemes belong to C_1, the cost is doubled because the number of morae is changed.

[3] Although the cost to compute the distance between all word pairs is high, it can be done beforehand.

[4] A mora usually consist of a consonant and a vowel. Since Japanese is a mora-timed language, the rhythm of Japanese places a beat on each mora in words, which is called mora isochronism.

Table 1: Examples of phonologically similar word pairs.

kyotofuritsu mogakko (京都府立盲学校) (Kyoto prefectural school for the blind) *kyotofuritsu rogakko* (京都府立聾学校) (Kyoto prefectural school for the deaf)	0.34
rakusai koko mae (洛西高校前) (In front of Rakusai high school) *rakusui koko mae* (洛水高校前) (In front of Rakusui high school)	0.36
kinkaku-ji michi (金閣寺道) (Kinkaku-ji temple road) *ginkaku-ji michi* (銀閣寺道) (Ginkaku-ji temple road)	0.37

The system extracts word pairs whose phonetic distances are low as phonologically similar words. Examples of such word pairs in the Kyoto City Bus Information System are shown in Table 1.

3.2 Obtaining Candidates for Clarifying Expressions

We describe how to obtain distinctive words as candidates for clarifying expressions to distinguish phonologically similar word pairs, which was described in the previous section. We focus on differences of phones or spelling differences between the pairs. The system obtains distinctive words that include the different spelling or phones from multiple knowledge sources. We used four the knowledge sources described later.

First, the system extracts characters, which are Chinese characters here, or phones that are not common between confusing word pairs. Hereafter, the first character or phone extracted is denoted as `<char>` or `<phone>` [6]. Their positions in confusing words are denoted as `<charLoc>` and `<phoneLoc>` represented using common characters (or phones) before the difference. An example is shown below for *rakusai koko mae* (洛西高校前) and *rakusui koko mae* (洛水高校前).

$$\begin{cases} \texttt{<char>} & = \text{“西”, “水”} \\ \texttt{<phone>} & = \text{“\textit{sa}”, “\textit{su}”} \\ \texttt{<charLoc>} & = \text{“after the character \textit{raku} (洛)”} \\ \texttt{<phoneLoc>} & = \text{“after the phones \textit{raku}”} \end{cases} \qquad (2)$$

When the difference is at the beginning (or end) of a word, we use `<charLoc>` = `<phoneLoc>` = “at the beginning”

[5] The edit distance of two strings, $s1$ and $s2$, is defined as the minimum number of mutations to change $s1$ into $s2$, where a mutation is a substitution, insertion, or deletion.

[6] Phones in Japanese correspond to the Japanese syllabary.

Table 2: Specifications of JC1.

difficulty	# words	example
A1 (low)	2,337	反対 (objection)
A2	2,618	主要 (main)
B	4,892	警告 (admonition)
C	8,989	歓待 (hospitality)
F (high)	16,239	違憲 (unconstitutionality)

合図 (sign), 金曜 (Friday), 空気 (air),
水曜日 (Wednesday), 反対 (objection), ...

Figure 4: Examples of pairs of Chinese characters.

Chinese character	native Japanese readings
外 (out)	*soto, hazusu, hoka*
音 (sound)	*oto, ne*
声 (voice)	*koe, kowa*
読 (read)	*yomu*

Figure 5: Examples of Japanese readings of Chinese characters.

(or "at the end").

The system obtains distinctive words from knowledge sources as candidates that contain the different spelling or phones. We will describe four methods corresponding to the knowledge sources used.

Method Using Words Consisting of Several Chinese Characters

The system uses words consisting of several Chinese characters including differences in spelling. We used a dictionary, JC1 [Sato, 2004], in which the difficulty rankings are given. The specifications of JC1 are listed in Table 2. We used words having the lowest difficulty to improve the intelligibility of expressions. This dictionary includes 2,337 words with the lowest difficulty, part of which is shown in Figure 4. The template to generate expressions of confirmation is as follows:

> "Does the character <char> in <ccpair> follow <charLoc>?"

Here, <ccpair> denotes a word in the dictionary. Inappropriate words are removed from the candidates; for example, when a word extracted from the dictionary is a part of its original confusing word pairs, etc.

Method Using Readings of Chinese Characters

We used Japanese readings of Chinese characters to explain confusing words. We adopted Japanese readings if the target Chinese character had other Japanese readings. Examples of the set of readings are shown in Figure 5. The template to generate clarifying expressions of confirmation is as follows:

> "Does the character read as <reading> follow <charLoc>?"

Here, <reading> denotes another reading of the target Chinese character.

Chinese character	English translation
雨	rain
黄	yellow
石	stone
橋	bridge

Figure 6: Examples of translations of Chinese characters.

phone	description word	
sa	*sakura*	(cherry blossom)
shi	*shimbun*	(newspaper)
fu	*fujisan*	(Mt. Fuji)

Figure 7: Part of Japanese phonetic code.

Method Using English Translation of Chinese Characters

We used English translation of a Chinese character when a single Chinese character had a meaning by itself. If a Japanese-English dictionary has an index entry that consists of a single Chinese character, we assume the entry has a meaning itself, and it is used with its English translation. We obtained 148 single Chinese characters and their translations, after limiting them to those with the lowest difficulty. Examples of the set of single Chinese characters and their translations are shown in Figure 6. The template to generate expressions of confirmation is as follows:

> "Does the character <char>, which means <trans> in English, follow <charLoc>?"

Here, <trans> denotes a translation of a target character.

Method Based on Phonetic Code

We also used the Japanese phonetic code to explain phones at the beginnings of differences. A part of the table is shown in Figure 7. We used a word in the table as a distinctive one to explain a phone. The template to generate expressions of confirmation is as follows:

> "Does the phone <phone> of <descword> follow <phoneLoc>?"

Here, <descword> denotes a word describing a phone in the Japanese phonetic code.

3.3 Selecting Best Candidates for Clarifying Expressions

We describe how to select the most intelligible expressions from candidates. As described in Section 2.3, intelligibility depends on both the difficulty of words and the ease in hearing. To reduce the difficulty, we only used words having the lowest difficulty rankings when obtaining candidates from dictionaries, as described in the previous section.

The ease in hearing is defined by the uniqueness of the distinctive word in the clarifying expression and the difference between the distinctive and original words. Here, we define the uniqueness, u_i, and the difference, d_i, of distinctive word w_i.

The uniqueness, u_i, represents how words are included that are phonologically similar to word w_i in the system's

Table 3: Average number of candidates and percentage of times that at least one candidate was obtained from sources (bus system).

average # candidates	percentage of times candidate was obtained			
	I	II	III	IV
3.9	44%	46%	13%	100%

I using words with several Chinese characters
II using readings of Chinese characters
III using translation of Chinese characters
IV using phonetic code

Table 4: Distribution of subjective labels of clarifying expressions (bus system).

	A	B	C
labels for all obtained candidates	113	102	13
labels for selected candidates	49	9	0

knowledge sources, such as sets of words consisting of several Chinese characters, a Japanese-English dictionary, and the Japanese phonetic code. The user will not confuse the word with others when few similar words are in the knowledge sources.

The uniqueness, u_i, of word w_i is defined as follows:

$$u_i = \frac{1}{N} \sum_{k=1}^{N} p.d.(w_i, \tilde{w}_k),$$

where $p.d.(w_i, \tilde{w}_k)$ denotes the phonetic distance defined in Section 3.1. Words having the k smallest $p.d.(w_i, \tilde{w}_k)$ in all knowledge sources are denoted as $\tilde{w}_k$. That is, u_i is defined as the average of the N smallest phonetic distances of words. Here, we set $N = 10$.

The difference represents how a distinctive word to be used in a clarifying expression differs from the original one. The clarifying expression is needed when the original expression is phonologically confusing. Therefore, it may still be confusing if the distinctive word is similar to the original one. The difference represents the intelligibility when original words are presented.

The definition of the difference, d_i, of distinctive word w_i is as follows. Here, w_{long} and w_{short} denote the words with more and less phonemes of the original word, w, and distinctive word w_i. The number of phonemes of w_{long} and w_{short} are N_l and N_s. The difference, d_i, is defined as the minimum of the phonetic distance between w_{short} and all partial phoneme sequences in w_{long} whose lengths are N_s:

$$d_i = \min_{k=0,..,N_l - N_s} \{p.d.(w_{long}[k+1 \,..\, k+N_s], w_{short})\},$$

where $w[k \,..\, k']$ denotes a partial sequence in w_{long} from the k-th phoneme to the k'-th phoneme.

Using the above two criteria, we assume that the more intelligible candidates have larger values of uniqueness and difference. We define a score representing the intelligibility of the clarifying expression including distinctive word w_i as the weighted sum of the uniqueness and the difference. The system compares the scores and selects the candidate that has the largest one.

$$score(w_i) = W_1 \cdot u_i + W_2 \cdot d_i$$

Table 5: Average number of candidates and percentage of times that at least one candidate was obtained from sources (hotel system).

average # candidates	percentage of times candidate was obtained			
	I	II	III	IV
3.3	40%	48%	15%	96%

Table 6: Distribution of subjective labels of clarifying expressions (hotel system).

	A	B	C
labels for all obtained candidates	102	60	9
labels for selected candidates	47	5	0

We set the weights as $W_1 = 4$ and $W_2 = 1$ after examining various values in a preliminary experiment [7].

4 Experimental Evaluation

4.1 Implementation into Existing Systems

We applied our method to the vocabulary of the Kyoto City Bus Information System [Komatani *et al.*, 2003]. The vocabulary includes the names of bus stops, bus route numbers, and names of famous places or public facilities near bus stops. The vocabulary size is 1,574.

Table 3 shows the average number of candidates per confusing pair and the percentage of times that at least one candidate was obtained from knowledge sources, when our method was applied to the vocabulary of the bus system. Candidates for clarifying expressions were not always obtained because each knowledge source may not a word including different characters or phones between confusing word pairs. The method using the phonetic code can generate candidates for all words because it focuses on phones and distinctive words are prepared for all phones in the Japanese phonetic code. As a result, on average, 3.9 candidates were obtained per confusing pair from the knowledge sources.

We evaluated whether the generated clarifying expressions were intelligible or not. We gave labels subjectively for each expression: from A (intelligible) to C (unintelligible). The distributions of the subjective labels for candidate expressions and those of the selected expressions with maximum scores are shown in Table 4. This table indicates that expressions with better subjective labels were mostly chosen.

We also applied our method to another system: a hotel search system [Komatani *et al.*, 2001]. The vocabulary includes 865 names: those of wider areas than the bus system uses, facilities of hotels, etc. In this domain, candidates were generated from each knowledge source, and 3.3 were generated on average as shown in Table 5. The selection module also functions well without modifications, as shown in Table 6.

4.2 Listening Experiment for Generated Confirmations

We also evaluated whether generated confirmations actually improved the ease in hearing the difference between phono-

[7] Although we set the values based on the bus system's vocabulary, it worked well in the hotel system described later.

Table 7: Results of conventional confirmation.

Subject	Correct	Incorrect	Cannot discern
A	21	0	34
B	38	8	9
C	23	1	31
D	25	0	30
E	32	0	23
Average	27.8 (51%)	1.8 (3%)	25.4 (46%)

Table 8: Results of confirmation using clarifying expressions.

Subject	Total	Correct	Incorrect	Cannot discern
A	34	26	0	8
B	9	5	2	2
C	31	14	3	14
D	30	21	0	9
E	23	22	1	0
Average	25.4	17.6 (69%)	1.2 (5%)	6.6 (26%)

logically confusing word pairs. The procedure of our experiment is as follows. Here, we denote each phonologically similar word pair as $word_1$ and $word_2$. First, a user was displayed $word_1$ as text. Second, he/she listened to the following confirmation generated by a TTS system, and answered "yes", "no", or "I cannot tell".

> I am not sure whether you said $word_1$ or $word_2$.
> Did you say X?

where X was either $word_1$ or $word_2$, determined at random. Then, if his/her answer was "I cannot tell", he/she also listened to the following confirmation generated by a TTS system, and answered similarly:

> I am not sure whether you said $word_1$ or $word_2$.
> *(confirmation with clarifying expressions for X)*

He/She could listen to each confirmation up to three times.

We added background noise to the TTS sound, because we assume that the system will be used in noisy environments. The volume of noise was set to equal that of the TTS sound, and we used background noises from shopping malls. Five subjects listened to 55 confusing word pairs that had been automatically extracted from the bus and hotel system vocabularies.

The results of the first confirmation for confusing word pairs "Did you say X?" are shown in Table 7. Almost half the word pairs could not be discerned although users were allowed to listen up to three times. Word pairs that could not be discerned in the first confirmation were confirmed again using the clarifying expressions. The results are in Table 8 show that 69% of the word pairs that could not be discerned in the first confirmation became discernible using the confirmations with clarifying expressions. That is, our method improved the intelligibility of confirmations for phonologically similar word pairs that could not be discerned using conventional confirmations. However, 26% of the word pairs still remained unintelligible after the confirmations using clarifying expressions. One of the reasons for this is that the distinctive words used in the clarifying expressions semantically had nothing to do with the context and could not be expected by users.

5 Conclusion

We described a method of generating clarifying expressions to prevent confusion between phonologically confusing word pairs in system vocabularies. A framework of our method can be applied to various domains because confusing pairs are extracted automatically and candidates for clarifying expressions are automatically generated and selected. Therefore, we can generate clarifying expressions without preparing them for each domain. We experimentally verified that clarifying expressions could improve listener abilities to discriminate between phonologically similar words.

Because our target language was Japanese, we adopted Chinese characters as units and generated clarifying expressions to explain them. If appropriate morphemes can be defined for other languages, phonologically confusing word pairs can be distinguished by preparing knowledge sources including the morphemes. This would be helpful even when users cannot accurately spell words.

The intelligibility of clarifying expressions also depends on user characteristics and the context. The distinctive words may be unrelated to the contexts. The intelligibility will improve further if such words were more familiar, which depends on individual users. Confirmations for the remaining 31% of words that could not be discerned in our experiment would be more intelligible if clarifying expressions were generated adaptively for individual users.

References

[Hazen *et al.*, 2000] T. J. Hazen, T. Burianek, J. Polifroni, and S. Seneff. Integrating recognition confidence scoring with language understanding and dialogue modeling. In *Proc. ICSLP*, 2000.

[Hirsch and Pearce, 2000] H.-G. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluation of speech recognition under noisy conditions. In *Proceedings of the ISCA ITRW ASR*, 2000.

[Komatani *et al.*, 2001] K. Komatani, K. Tanaka, H. Kashima, and T. Kawahara. Domain-independent spoken dialogue platform using key-phrase spotting based on combined language model. In *Proc. EUROSPEECH*, pages 1319–1322, 2001.

[Komatani *et al.*, 2003] K. Komatani, S. Ueno, T. Kawahara, and H. G. Okuno. Flexible guidance generation using user model in spoken dialogue systems. In *Proc. ACL*, pages 256–263, 2003.

[Navarro, 2001] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[Raux and Eskenazi, 2004] A. Raux and M. Eskenazi. Non-native users in the let's go!! spoken dialogue system: Dealing with linguistic mismatch. In *Proc. of HLT/NAACL*, 2004.

[Sato, 2004] S. Sato. Identifying spelling variations of Japanese words (in Japanese). In *IPSJ SIG Technical Report*, 2004-NL-161, pages 97–104, 2004.