인지적 철자 교정 후보 제시기:
삽입, 생략, 전위, 대치 오류 수정을 위한 복합 방안

이 종호, 이 종혁, 이 근배
포항공과대학교 전자계산학과

COGNITIVE SPELLING THERAPIST:
A combined method
for correcting four types of spelling errors:
insertion, deletion, transposition, substitution

JONG-HO LEA, JONG-HYEOK LEE, & GUNBAE LEE

Pohang University of Science and Technology
Department of Computer Science
San 31 Hyoja-dong
Pohang 790-784 Republic of Korea

Cognitive Spelling Therapist generates the candidates for correction of one-letter misspelling words, which correspond to over 80 % of the misspelling words. One-letter misspelling can be divided into four categories, and for each categories Cognitive Spelling Therapist copes them with seperate cognitive therapies. Each therapy is based on cognitive causes of misspelling: figural confusion, pronunciation confusion, and keyboard confusion. Cognitive Spelling Therapist generates three candidates for correction in average. After we tested the correctness of candidates with 185 misspelled words randomly sampled from two typist for two months, Cognitive Spell Therapist showed 97.5 % correction for substitution errors, while insertion, deletion, and transposition errors were perfectly corrected.

Nowadays one of the most popular way to make a file is using the word processor in a computer. The word processor uses the keyboard as a main input device. The typist should switch his/her attention from the text on the bookrest(or on his/her mind) to the monitor screen, and even to the keyboard. Switching attention between text, screen, and keyboard makes the typeist to slip during his/her typing. Even though the typist is the specialist, typing errors occur and survive against proof reading. If typing errors can be detected and corrected by the corrector automatically, we may relieve ourselves from the suspicion of the existence of the typing error in the text..

In this thesis, we intend to present a new combined way to correct a misspelled word. This thesis is the part of the big project that we would continue to work out(Fig. 1). Our corrector, Cognitive Spelling Therapist, assumes that the misspelled word has already detected, and that candidates of corrections will be selected into one correct word by context-selection mechanism.

First, we would look at the specification of typing errors.

1. Specifications of typing errors

The first dimension of typing errors, which researchers have considered on, is the number of misspelled letters in a word. Damera (1964) and Peterson(1980) had shown that almost 80 % of spelling errors is from the misspelling of only one letter.

The second dimension is the causes of spelling errors. There are two kinds of errors(Peterson, 1980). One is from ignorance of the speller (they call it orthographical error). The other is from errors in movements of fingers(i.e. typographical error).

The third dimension is persistence of error occurrence(Peterson,

본 논문은 과학재단(KOSEF)의 연구과제 92-21-00-05 "필기체 한글을 위한 후처리 기법에 관한 연구"의 오인식 처리 관련 연구임.

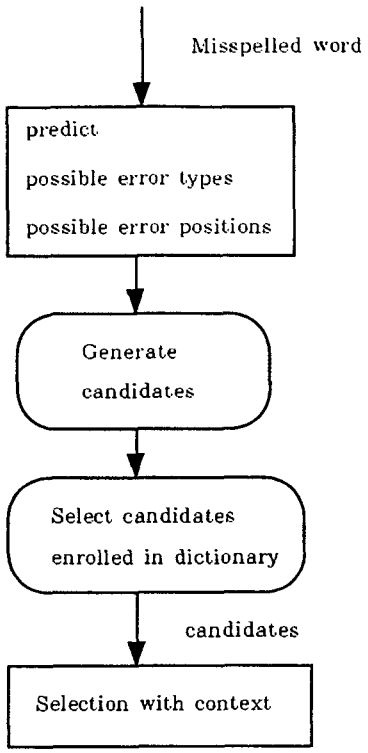

Fig 1. Dataflow of candidate generator

1980). Some researchers believe that orthographical errors (which occur when a speller does not know the correct spelling) is more persistent in occurrence than typographical errors (which occur when a speller hits the wrong key on keyboard). And they believe those orthographical errors usually come from phonological confusions of the spellings.

The fourth dimension is types of forms of spelling errors. Damerau (1964) and Peterson(1980) specified four categories. One is the transposition of two adjacent letters in a misspelled word. Second, there is usually one extra letter in a misspelled word. Third, one letter in a word is missing. Fourth, one letter in a misspelled word is wrong one.

2. Some existing correction methods

It will be helpful for understanding the spelling errors to look into the existing correction methods. They could be divided into three main categories in accordance with their methods. One uses statistical information of the text which contains the spelling errors(Pullock and Zamora, 1984), or of the groups of letters(i.e. trigram) in the words(De Heer, 1982; Angell, 1983). The second uses specific linguistic information(phonological rules) for correcting the misspelled errors(Van Berkel, 1986; Daelemans, 1987; Berkel and Smedt, 1988). The third uses a dictionary to find out the most similar word in the dictionary by comparing with the misspelled word(Hall and Dowling, 1980).

These methods strongly rely on the statistical methodology--- even Berkel and Smedt(1988) tried to combine phonological rules with trigram analysis. In contrast, we would rather put emphases on lack of attention as causes of typing errors. Lack of attention causes typing errors, and leaves several cues in a misspelled word just like a footprint. Therefore, we may use these cues in the misspelled word itself(rather than using the statistitical markov assumption) for unraveling the puzzles---correcting the typing errors.

3. Review of Causes of typing errors

Once we have a look at the cognitive pathway of the typing behavior, we may easily find where the errors might happen by what reason(Fig. 2).

When a speller is reading a note to be typed, he/she looks into the individual spellings of the words. At the same time, the speller is rehearsing the letters with their pronunciations while he/she is typing. By rehearsing, he/she puts the words verbatim into his/her short term memory(STM) by pronunciations of the words, and gets their meanings. At that time, he/she may be enticed by the pronunciation of the words(because of the prevalence of auditory information in STM), and may type the words not exactly as they look but just as the words are pronounced. Similarily, because the speller may put his/her attention to the meaning of the sentence, he/she may not take the figures of the letters seriously. Then, he/she may type another letters with similar figure instead for the letter to be typed. Otherwise, he/she may conserve the spelling in his/her mind, but his/her fingers may not

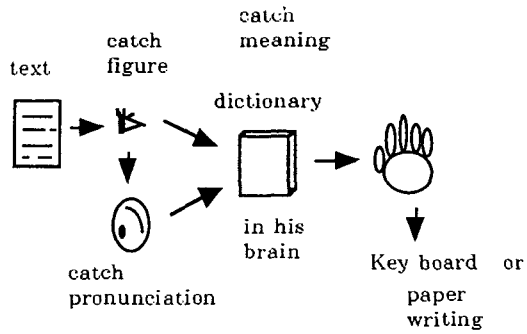

Fig 2. Cognitive pathways of reproducing spelling of words

follow the spelling in his/her mind correctly. Then, he/she may punch an adjacent key, instead.

Correspondingly, we may review the two causes of typing errors(orthographical errors and typographical errors) with this cognitive view. Orthographical errors are from uncertainties of the correct spellings or from uncertainties of rules of pronunciation(which match between the letters and pronunciations). In this situation, spellers are trying to use pronunciation of the word as the basis for spelling, because they do not know the correct spelling. But, there are lots of mismatch between pronunciation and spelling, so these mismatch would cause orthographical spelling errors. Typographical errors are from errors in finger movements, but sometimes fingers do not have the whole responsibility of the typing errors. Cognition(perception and meaning processing) may be distorted from incorrect reading, incorrect pronunciation, and incorrect interpretation of the word. Therefore, orthographical errors and typographical errors have several causes that have to be more specified with cognitive view.

Eventually, we may understand the causes of typing errors as several kinds of confusions that take place along the cognitive pathway and are resulted from the lack of attention. Along the cognitive pathway, there are three major causes that induce the confusions. First, incorrect reading of the spelling takes place when we read the text to be typed. Figural confusion between the letters leads the typist to make mistakes, even though the typist knows the correct spelling. Second, Mismatch between spelling and pronunciation requires the typist to pay more attention to spelling, but many times typist has to rush their typing task, so he/she just types the pronunciation of the letters instead of the correct spelling. Sometimes the mismatch between spelling and pronunciation can be resolved by knowing of orthographical rules of the language(i.e. avoiding orthographical confusion), whereas pronunciation confusion like that between 'b' and 'p' is hard to escape(i. e. pronunciation confusion). Third, Inexactness of keyboard punching is caused by the confusion of the positions of the keys on the keyboard.

On the whole, the first cause relates to the perception process of the cognitive pathway of typing, while the second cause does to the rehearsing in STM. The third cause relates to the action of the fingers, even the typist has correct perception, rehearsing, and orthographical knowledge.

As we wish to find out the relationships between the causes of typing errors and the types of typing errors, in this thesis we narrow down the number of misspelled letters in a word to only one letter misspelling------Damerau(1964) and Peterson(1980) said that almost 80 percent of spelling errors is one-letter misspelling, and our data of misspelling words(n= 185) contains 169 one-letter misspelled words(91.4%).

One-letter misspelling can be divided into four classes: insertion, substitution, deletion, and transposition. Insertion is an error when a speller puts an additional letter in a word (e.g. inseertion, instead of insertion; 삽ㅇ입 for 삽입). Substitution occurs when a letter in the word is substituted by other letter (e.g. sutstitution, instead of substitution; 대채 for 대체). Deletion is an error which a word has not a letter for its full spelling (e.g. deleton, instead of deletion; 새략 for 생략). Transpositon occurs when two adjacent letters in a word change their positions together (e.g. tranpsositon, instead of transposition; 정뉘 for 전위). Although transposition has two misplaced letters in a word, substitution is one-letter misspelling. If we put the first misplaced letter back in its right place without moving the second one, the misspelling will be corrected.

We randomly collected 92 one-letter misspelled Korean words from two professional typists for two months, and enumerate them with regard to four error types and error positions in the words(Table 1). Table 1 shows that substitution error is the most frequent one(71.7%), and that deletion(14.1%), insertion(11.9%), and transposition(2.3%) follow substitution in the number of occurrence. Transposition has rarely found in our data(only 2 out of 92) than we expected, and we may attribute this result to the figural constraints in the rule of writing Korean syllables.

Korean has a formal rule for writing syllables: each syllable is seperated from other syllables in figure, which contains three sounds. The first sound of a syllable is called "cho-seong" . And the second and the third are called "joong-seong" and "jong-seong" , seperately. These three sounds has their own position in a syllable. The first sound, "cho-seong" , locates at left-upper part of syllable. The second sound, "joong-seong" , locates at right-upper part of syllable, while the third sound, "jong-seong" , locates at lower part of syllable, and it could be optional in some syllables. For example, if we show this rule by using English word "pendulum", we can seperate this word into three syllables: "pen-du-lum°(펜-두-럼). The first sound, "cho-seong" , of the second syllable is "d" (ㄷ). The second syllable, "du" (두), has only two sounds. The third sound, "jong-seong" , of "du" (두) has no sound value, and in only that case we do not mark it in a syllable.

These figural constraints in Korean syllables make any transposition more salient than normal syllables, therefore, typist can detect his/her transposition errors more easily than English words which contain transpositions.

In this paper, we considered words which have less than four syllables(because most Korean words are within the range of below four syllables). We can see that the second syllable is the most

Table 1. Distribution of error types and error positions

error types	positions of syllables												
	1st syl			2nd syl			3rd syl			4th syl			
	1	2	3*	1	2	3	1	2	3	1	2	3	sum
subst	10	8	4	13	5	8	6	6	3	1	2		66
del			3			1 4	1	2	1			1	13
ins						3 3			3			2	11
trans			2										2
total	25			39			22			6			92

* 1 2 3 stands for cho-seong, joong-seong, jongsoeng.

probable place where misspelling may occur. The first and third syllable are the next to top. The fourth syllable is the least probable place where error may happen(Table 1).

Our analyses of 92 words suggest that all four types of spelling errors may have relationships with three causes that we have seen before: keyboard, figure, and pronunciation confusions. As a prototype, susbtitution errors may give you a clear view of the relationships with three causes(Table 2). Errors from keyboard-position can explain a little more than half(52%) of substitution errors, but not whole. Errors from figure confusion(27%) occur similarly in proportion to errors from pronunciation confusion(21%).

Table 2. Ramification of substitution errors

Causes	positions of syllables: 1st syl	2nd syl	3rd syl	4th syl	sum
Keyboard					34
up	1	3		1	(52%)
left	2	8	4		
right	3	2	1		
down	6		1		
without shift-key		1			
adding shift-key	1				
Figure	7	6	4	1	18 (27%)
Pronunciation	2	6	5	1	14 (21%)
total	22	26	15	3	66

4. Cognitive Spelling Therapist for correcting misspelled words

Because misspelling in a word occurs almost free of context, it is hard to decide what type of errors happens and where it locates. Even though we can not tell what types of error happen at the misspelling words, we can constrain the scope of possibilities of misspelling by considering the cognitive factors, and find the therapies of these error types from these possibilities. This is why we call our candidate generator of misspelled word as Cognitive Spelling Therapist.

Cognitive Spelling Therapist has seperate therapies for correction of four error types, seperately(Fig 3.). Substitution errors are replaced by the proxies of replacements. Deletion errors are filled by fill-in candidates. Culling an extra letter in a misspelled word corrects insertion errors. Transposition errors are corrected by transmigration of the missplaced letters. Proxies for substitution and fill-in candidates for deletion are derived from three causes at the cognitive pathway.

5. Architecture of candidate generator

(Cognitive Spell Therapist)

We just concentrate on the misspelled word itself as the basis for correction. We do not use the frequencies of letters in the text for statistical comparison just like trigram analysis.

We predict the possible types of errors, and then correct the misspelled word by the predicted types. The typing errors happen from fatigue or lack of attention at that moment. It is very hard, however, for us to detect the exact position where the typing errors occur.

On the basis of four types of misspelling and three causes of typing errors, we can suggest the candidates of corrected words with respect to four types of error. For this purpose, we need four packages of rules for the four types of error. These are "tables for proxies" (i.e. tables for correction of substitution errors), "fill-in rules" (i.e. rules for correction of deletion errors), "culling rules" (i.e. rules for correction of insertion errors), and "transmigration rules" (i.e. rules for correction of transposition errors).

Tables of proxies are divided into four tables. These are "orthographical law table", "figure confusion table",

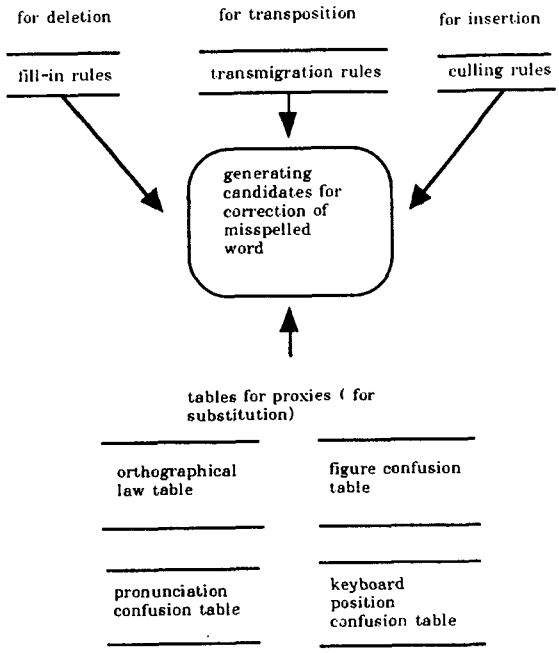

Fig 3. Architecture of Cognitive Spell therapist

"pronunciation confusion table" , and "keyboard position confusion table" . These four tables are used to provide the scope of proxies of letters. So these tables are called as "tables for proxies" . First, a table for correcting the orthographical errors is needed. This table is named "orthographical law table" . Second, we need a table for correcting the substitution error caused by similar figure of letters. We call this table as "figure confusion table" . Third, there are scopes of possible letters that a speller may mistakes them by their pronunciations. We gather these possible letters into "pronunciation confusion table" . Fourth, false finger-movement has a limit such that we assemble the possible keys that a typist might hit improperly. These possible keys are gathered into "keyboard position confusion table" .

Orthographical law table contains phonological and orthographic rules, and proxies for letters in compliance with the phonological rules(Fig 4). In Korean, there are several condensed rules about relations between phonology and spelling(Korean has institutionalized orthographical law; they call it "maz-tchum-beob" : 맞춤법, 표준 발음법), so we can use these rules in a table form. For example, in Korean there is a law, named "law of the first sound(두음 법칙)" , which controls the combination between the first consonant and the second vowel in the first syllable. The first consonant, "n(ㄴ)" , can not combine with the second vowel like "yeo(ㅕ)" , "yo(ㅛ)" , "yu(ㅠ)" , or "i(ㅣ)" . So, the first consonant, "n(ㄴ)" has to be changed into "o(ㅇ)" in this case. In English, the fact that the consonant "q" has to be followed by the second vowel "u" is similar with "law of the first sound(두음 법칙)" , but not so much.

Figure confusion table is comprised of proxies of letters with regard to similarities of figures(Fig 5). Similarities of figures are decided by several criteria. These are the existence of corners or circles in the figure of letter, the frequency of them(if any), the degree of angle at the corner of the figure, and closedness of the form(Kim & Kim, 1992) . Because a speller does not confuse easily the correct letter with ones with dissimilar figure, there could be a scope for candidates of similar letters. We can decrease the size of candidates by computing the similarities from the given misspelled letters, instead of trying the whole possible alphabets as candidates. Thus, we can decrease the backtracking space to almost a fifth of the whole alphabet size. For example "ㅂ" and "ㅁ" are too similar to differentiate each other as "b" and "p" , or "m" and "n" are. Thus, the misspelled word "baint(베인트)" can be changed into "paint(페인트)" by Cognitive Spelling Therapist.

IF

focus syllable	focus position	focus letter	the second vowels
1	1	n(ㄴ)	yeo(ㅕ) yo(ㅛ) yu(ㅠ) i(ㅣ)

THEN

changing letter
o(ㅇ)

Fig 4. Orthograpical law table

Pronunciation confusion table is comprised of proxies of letters with respect to similarities of pronunciation(Fig 6.). The similarities of pronunciation can be measured by two different ways: one for consonant and another for vowels.

Because Korean consonants are made after the forms of the organs of pronunciation(such as mouth, tongue, and throat), they have close

relationships with the orgrans of pronunciation. We can use the position of organs where the sounds are made and the way of articulation as the criteria of pronunciation similarity of consonants (Cho, 1985). The way of articulation can be divided into six categories: stops, fricatives, affricates, nasals, laterals, and semivowels. For example, stops explain the similarity between "p(ㅍ)" and "t(ㅌ)". Stops mean when you sound "p(ㅍ)" you can feel stopping of articulation at the end of the sound "p(ㅍ)". The organ positions of pronunciation can be divided into ten categories: bilabial, labia-dentral, dental, alveolar, palato-alveolar,

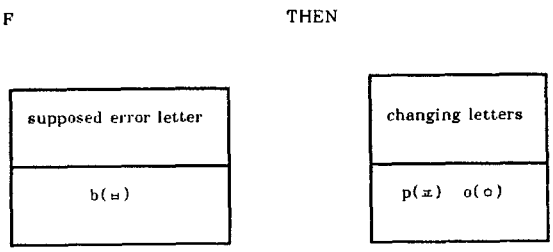

Fig 5. Figure confusion table

palatal, velar, uvular, pharyngeal, etc. For example, bilabial position means the lateral side of your lips, so "b(ㅂ)" and "m(ㅁ)" are similar because they use the lateral side of lips together.

Likewise, we use position of tongue and figure of lips as the criteria of similarity of Korean vowels(Cho, 1985). For example, "u(ㅜ)" and "o(ㅗ)" are similar because lips protrude and tongue locates higher than middle position. Thus, the misspelled word "tungue(퉁)" will be corrected into "tongue(통)". Also, this table can reduce the possible alphabets for correction to a fifth of the whole alphabet.

Keyboard-position confusion table contains proxies of letters with regard to the adjacencies of keys on keyboard(Fig 7.). The mistakes of finger movement are not too inaccurate for us to suggest the candidates. The collected data shows that a finger has a restrictive scope of making mistakes for each letter. From the 66 words which are classified as substitution errors, we found that keyboard position errors occur just within the range of four adjacent keys to the correct one. Bailey(1989) has shown that the left and the right key to the correct one are the most probable keys which a speller may make a mistake. Choi, et al.(1992) proposed that all adjacent keys(up to 10 keys) to the correct one be checked for correction. Our data, however, shows that only four adjacent keys are necessary for correction of key-confusion: left, right, up, and down key. Left, right, up, and down show the positions of misspelled key from the correct one. For example, "d" key in qwerty keyboard can be mistyped by only "s" (left), "f" (right), "e" (up), and "c" (down) key. In our data, spellers did not mistakes "d" for "w" (upper-left), "r" (upper-right), "x" (lower-left), or "v" (lower-right) as Choi et al.(1992) suggested.

Deletion errors can be corrected by "fill-in" the lost letter in the right place. Fill-in rules are comprised of the possible conditions in which the proxies can be filled, and of proxies of the deleted letters. For example, "biliards(다구)" will be corrected into "billiards(당구)".

Culling rules select and delete an unnecessary letter in a misspelled word. This rule just deletes one letter from the first letter step by step. Therefore, the insertion error like "rolle(역활)" will be changed into "role(역할)".

Transmigration rules are used for correction of transposition errors. Because each syllables has only three letters in Korean and vowel can locates only at the second position in each syllable, the transmigration rules are rather complicate with these positional constraints. Sometimes transposition causes to make additional syllable which does not exist in the correct word, or vice versa. For example, "em-ply-oee(직우너)" will be changed into "em-ployee(직원)"

6. Algorithm of candidate generator

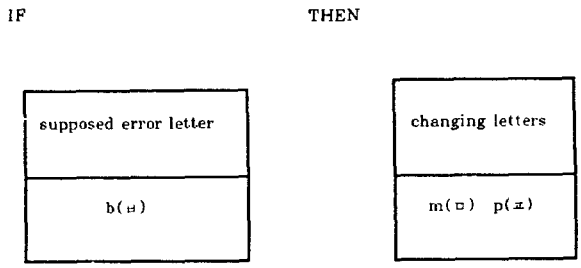

Fig. 6 Pronunciation confusion table

(Cognitive Spelling Therapist)

First, we check whether the word is in the dictionary. The words which are correct in spelling but not fit in context are out of interest in this paper. Second, if the word is not enrolled in the dictionary, Cognitive Spelling Therapist tries to generate the candidates of it for each type of errors. Third, we check whether the candidates are in the

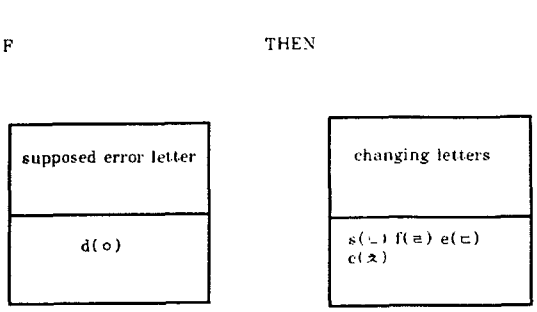

Fig 7. Keyboard position confusion table

dictionary, and report only the enrolled candidates.

At first, Cognitive Spelling Therapist checks whether the misspelled word has an independent consonant or vowel which consists of a single syllable. If it does, Cognitive Spelling Therapist tries to transmigrate, to fill-in, and to cull for correction. Otherwise, Cognitive Spelling Therapist tries to transmigrate, to fill-in, to cull-in, and to find better proxies for substitution errors. After Cognitive Spelling Therapist generates all candidates, the dictionary will sift them so that the qualified candidates will survive for correction.

7. Results of correction by candidate generator

Cognitive Spelling Therapist uses its rules in proportion to the size of misspelled words. The number of rules for correction grows linearly as the size of misspelled words does(Fig 8.). The size of the misspelled word can be measured by the number of syllables of it. Even though larger misspelled word uses more rules for generating candidates, the number of candidates which are enrolled in the dictionary is always under 5(three on the average). And the correct one among candidates locates at almost the second place among three.

The correctness of Cognitive Spelling Therapist is 97.5%(Fig 9.). Almost three percent of failure in correction occurs only in substitution errors, whereas deletion, insertion, and transposition errors are corrected successfully. On average, the misspelled words below 5 syllables(or 15 letters) uses 46 rules, which generate 3 candidates enrolled in dictionary.

8. Discussions

Our data was limited under five-syllable words, so we need to gather more real data above five-syllable word and to test whether Cognitive Spelling Therapist can cope the long misspelled-words correctly.

The further research about the conditions of occurrence of each types of errors is needed. The research on fatigue of finger movements(Park, 1992) would be helpful. If the correct estimation of the type of errors is possible, the number of candidates can be smaller even if we do not use the context information.

And the research about the context of the sentence is necessary to select the correct word from several candidates which are enrolled in dictionary.

Because Cognitive Spelling Therapist can treat only one-letter errors in spelling(80% of whole spelling errors) successfully, the way of correction for the words which contain more than two-letter errors has not been solved yet. Two-letter errors can be treated by Viterbi algorithm or etc. Therefore, research for correcting more than two-letter errors is still needed.

Although some rules of Cognitive Spelling Therapist are specified to Korean language, others rules can be generalized to English spelling corrector. For example, Keyboard position confusion table can be used in English because it is based on QWERTY keyboard.

The distribution of spelling errors may be different between Korean and English. For example, English has no positional constraints in syllable, so speller may not detect the transposition errors so easily as Korean does. The percentage of transposition error in English may be greater than that in Korean. But, transposition errors in English may be coped easily by Cognitive Spelling Therapist.

The propensity to checking the spelling by a speller is the main factor of preventing the misspelling. We gathered 185 misspelled words from two typists for 2 months, and only 16 words were more-

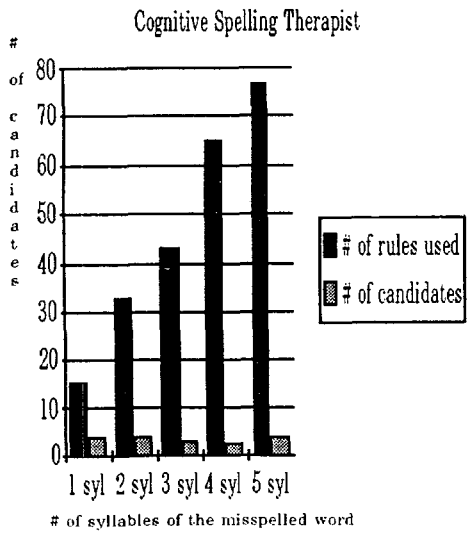

Fig 8. # of candidates with word size

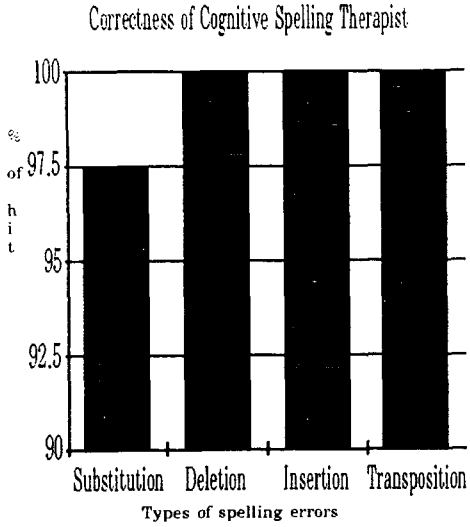

Fig 9. Test results of correctness(n = 169)

than-two-letter errors. It was only 8.6% of all, and most of them were retyping of the same syllable or misinterpretion of the figure of vowel as a part of consonant.

The systematic approach for gathering the real data of spelling errors will be very helpful for generation of more efficient heuristics for correction.

References

Angell, et al. (1983). Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19, 255-261.

Cho, Myeong-han. (1985). Psycholinguistics. Seoul: Min-eum-sa.
Choi, Ki-seon. & et al.(1992). The research on Korean spelling correction and parting of words. *The second year technical report.* KAIST.

Damera, F.J.(1964). A technique for computer detection and correction of spelling errors. Comm. ACM 7.3(March 1964). 171-176.

De Heer, T. (1982). The application of the concept of hoeosemy to natural language information retrieval. *Information Processing & Management.* 18, 229-236.

Daelemans, W. (1987). Studies in Language technology. Ph.D. Dissertation, Linguistic Dept., University of Leuven.

Hall and Dowling.(1980). Approximate String Matching. *ACM Computing Surveys*, vol. 12. December, 381-402.

Kim, Jeong-oh. & Kim, Jae-kab. (1992). Perception of letters and their processing during recognition of Korean words. *The fourth conference of Hangueol and Korean information processing.*

MunGyoBu.(1988). Hangueol maz-tchum-beob (Korean Orthographical law). The anouncement of Korean ministry of Education. No. 88-1.

Park, Hueong-oh. (1992). The basic research on Korean keyboard. In Lee, C.J. The basic research on Korean code and Korean keyboard. *Technical Report.* Hangueol and Computer.

Peterson, J.L.(1980). Computer Programs for Detection and Correcting Spelling Errors. *Communications of the ACM*, vol. 23, December 1980, 676-687.

Pullock, J.J. & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *CACM*, 27, 358-368.

Van Berkel, B. (1986). SPELTERAPUIT: een algoritme voor spel- en typefoutencorrectie gebaseerd op grafeem-foneemomzetting. Master°Øs thesis, Dept. of Psychology, University of Mijmegen.

Van Berkel, B. & Smedt, K. D. (1988). Triphone analysis: A combined method for the correction of orthographical and typographical errors. *The second conference on Applied natural language processing.* (Feb. 1988). Association for computational linguistics.