



Evaluating the Potential Effectiveness of Automatic Document Analysis

JAMES R. LEWIS

IBM Corporation, 8051 Congress Ave, Suite 2227, Boca Raton, FL 33487, USA

jimlewis@us.ibm.com

Abstract. This paper documents the motivation, method and results of seven experiments conducted to investigate the properties of automatic document analysis (for the purpose of automatic vocabulary expansion of a personalized language model in a speech dictation system). The results indicated that automatic document analysis of corrected text should improve the accuracy of text dictated in the future, as long as the future text is similar to the analyzed text. None of the manipulations had a measurable effect (either good or bad) when the analyzed text was uncorrected dictation or future text that was not similar to analyzed text. These results were the same for both trained and untrained acoustic models.

Keywords: automatic document analysis, automatic vocabulary expansion, dictation accuracy

Introduction

Recognition accuracy is an important attribute of speech dictation systems, but at all except the highest levels of recognition accuracy the speed of correction is as or more important for determining true throughput—the rate of production of corrected text (Lewis, 1999). In most speech dictation systems, the system only ‘learns’ from corrections that users make with the system’s correction dialog. A correction dialog is a component in the user interface that typically contains, for a selected misrecognition, a list of potential alternates and a free-form text entry field in which the user can type the correct text if it is not available in the list of alternates.

Careful examination of the correction behavior of people using the major commercial speech dictation products, however, has shown that users typically prefer to correct misrecognitions by typing directly into the text rather than using a correction dialog (Karat et al., 1999). An alternative way for a system to identify words that are out-of-vocabulary (OOV) and to add document-specific information to the language model (LM) is to run a program designed for the specific purpose of document analysis. The LM is an internal component of a speech dictation system derived from a computational analysis of a substantial body of text

data, modeling the probability of occurrence of a word given the preceding several words.

Most modern speech dictation systems (such as IBM ViaVoice[®], starting with ViaVoice ’98) provide a document analysis function (called, in ViaVoice ’98, Vocabulary Expansion). One purpose for this function is to let users analyze existing documents during installation for the purpose of improving the product’s recognition accuracy by identifying OOV words in a user’s selected documents and biasing the LM toward the user’s writing style. If it were possible to modify this function to allow automatic analysis of documents after a dictation session, then users might be able to make their corrections directly in the text of their dictated documents and still have the system learn about the changes. The system would acquire the same information as that formerly obtained through the correction dialog.

This paper describes seven experiments conducted during late 1998 and early 1999 to investigate the potential effects of automatic document analysis. In general, we wanted to understand these effects given variation in user enrollment, LM updating, and correction of misrecognitions. Enrollment, also referred to as training, is the process of getting samples of a user’s speech for the purpose of computing a personalized acoustic model. This is the defining feature of systems that are speaker dependent. Updating a LM can, at a minimum,

include the process of adding new words to the recognizer's vocabulary. It is also possible to update word n -gram likelihood probabilities (typically, unigram, bigram, and trigram). Variation in the correction of misrecognitions is important because users might or might not make these corrections before performing an action that would cause automatic document analysis to occur. Other issues of interest were the comparison of the effect of automatic document analysis with that of standard correction using a correction dialog and the effect of analyzing randomly generated text on future recognition accuracy. The specific experimental questions for the seven studies were:

- What if the automatic document analysis, run after correcting misrecognitions, only added new words, but did not update word likelihood probabilities in the user's personal LM (also known as the 'user cache') for users who have completed enrollment?
- What if the automatic document analysis, run under the conditions described above, also updated the word likelihood probabilities in the user's personal LM?
- How would automatic document analysis compare with the standard correction procedure for the purpose of improving future dictation accuracy for enrolled users?
- What if enrolled users didn't correct the text before automatic document analysis?
- What if unenrolled users didn't correct the text before automatic document analysis?
- What if unenrolled users did correct the text before automatic document analysis?
- What if an external audio source caused the accidental production of random text that the system automatically analyzed, resulting in contamination of the personal LM?

The decision to deploy development resources to the coding of an automatic document analysis feature for future versions of ViaVoice (those following ViaVoice '98) depended on the outcomes of these experiments.

General Method

Participants

The participants were four males and four females (all adult native speakers of American English), all of whom had provided recordings of test scripts (using

Sonic Factory's Sound Forge[®], 22 kHz, 16-bit mono PCM recording) and enrollments (speaker-dependent acoustic models) from a previous unpublished study of the effectiveness of ViaVoice '98 enrollment.

Materials and General Procedure

The computer used to process the recordings was an IBM Pentium[®] Pro 200 (48 MB memory, Windows 95, hundreds of MB of free hard disk space)—the same system on which participants had enrolled and recorded test scripts. The version of ViaVoice installed on the computer had a customized function designed to let an experimenter define a wave file for speech-to-text transcription.

The eight test scripts (selected from a small library of test scripts) included words, punctuation, and formatting commands. To create a set of texts related to another set, I divided the first four scripts approximately in half. The rationale for doing this was to define the first half as a treatment set and the second half as a related set (making the reasonable assumption that the two halves of a test script would have a stronger relationship with each other than with any other test script). The second four scripts contained material unrelated to the first four, and acted as a control set of texts. Table 1 shows the number of words and OOV measurements for each test script.

Before running the experiments described in this report, all recorded files were run through the customized transcription function to determine the baseline accuracy for each file (the four files that the experimenter would treat in the experiments, the four files of related text, and the four control files). In each experiment, the effect of the experimental treatment was assessed for both related and unrelated (control) texts by:

- subtracting the treatment word error rate from the baseline word error rate for each speaker,
- converting these differences to percentage reductions in word error rate by dividing the differences by the baseline word error rates,
- and computing a difference-score t -test for each set of data (related text and unrelated text).

I used two difference-score t -tests (also called paired observations t -tests, see Steele and Torrie, 1960) in Experiments 1–6 rather than analysis of variance because these planned t -tests provided very direct tests of the effect of the experimental variable, clearly parsed between the related and unrelated texts. Furthermore, the

Table 1. Word counts and out-of-vocabulary measurements for test scripts.

Script	Use	Words	# OOV	% OOV	OOV Words
Ford-1	Treated	193	0	0.0	<none>
Nick-1	Treated	134	3	2.2	Nickell, Roswell, lifecast
Laser-1	Treated	211	7	3.3	corneal (2), LVC, PRK, excimer, keratectomy, photorefractive
Tax-1	Treated	423	0	0.0	<none>
	Total	961	10	1.0	
	Mean	240.3	2.5	1.4	
Ford-2	Related	160	1	0.6	Gilbreth
Nick-2	Related	154	2	1.3	Roswell (2)
Laser-2	Related	190	0	0.0	<none>
Tax-2	Related	448	1	0.2	Noncharitable
	Total	952	4	0.4	
	Mean	238	1	0.5	
Bart	Control	276	3	1.1	Bartholomew, Kurtz, spate
Hoff	Control	303	0	0.0	<none>
Ad	Control	250	0	0.0	<none>
Health	Control	73	0	0.0	<none>
	Total	902	3	0.3	
	Mean	225.5	0.8	0.3	

entire set of experimental manipulations did not fit into a framework conducive to evaluation via analysis of variance. A subset of the experiments (Experiments 2, 4, 5 and 6) was suitable for evaluation via analysis of variance, with that analysis appearing in the Summary of Results section. Note that even with thirteen *t*-tests conducted at $\alpha = .05$, the number of tests expected to have a significant outcome by chance is less than one ($.05 \times 13 = .65$ —see Abelson, 1995 for additional arguments concerning the legitimate use of multiple *t*-tests).

Some Relevant System Characteristics

In general, the recognition methods used in ViaVoice are essentially the same as those used in other commercial speech dictation products. References describing these methods are available in Das and Picheny (1996) and Jelinek (1999). Note that although Das and Picheny were describing experiments conducted with

an isolated rather than a continuous speech recognition system, ViaVoice uses the same methods for adding words to the system vocabulary (M. Picheny, personal communication, September 11, 2002). The following characteristics are of particular interest (see the references for mathematical descriptions):

- The LM used an *n*-gram statistical model (M. Picheny, personal communication, March 26, 2003).
- The method used to produce acoustic models was fenonic modeling (Bahl et al., 1988). Fenonic modeling is a data-driven technique that typically outperforms manual phonetic transcription (Das and Picheny, 1996).
- The personal LM (cache model) allows adaptation of an otherwise static LM. In ViaVoice, the system constructs a dynamic trigram LM from the *m* most recent words. Because the next word might not be in the cache and the cache contains only a very limited set of trigrams, the cache model is generated from a linear interpolation of the dynamic model with the conventional static trigram model. See Das and Picheny (1996) for computational details. This cache model has the effect of biasing LM probabilities towards recently used text.
- The method for adding new words is an information theoretic procedure that uses both the spelling of the new word and examples of its pronunciation to produce at least one baseform (Bahl et al., 1991; Das and Picheny, 1996; Lucassen and Mercer, 1984).

Experiment 1: Normal vocabulary expansion without updating the user cache language model.

Motivation. The goal of this study was to characterize the effect of vocabulary expansion (finding and adding acoustic models for OOV words) but without updating the *n*-gram (unigram, bigram, and trigram) word probabilities in the language model for the speaker. This had the effect of isolating the influence of simply adding OOV words to the user cache and providing an estimate of the effect of this very conservative approach to vocabulary expansion.

Method. The strategy in this experiment was to run the treatment text (the first half of the first four documents) through the ViaVoice Vocabulary Expander after restoring a participant's enrollment (saved previously with no information in the user cache). To make it easier to run the text through Vocabulary Expander, I created a single file from the first half of the scripts in

Table 2. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 1.

Speaker	Related	Control
1	8.9	-20.4
2	11.9	-9.3
3	16.8	-4.7
4	6.2	3.1
5	13.4	-6.5
6	-3.1	8.7
7	3.1	-2.6
8	-1.5	-1.3
Mean	7.0	-4.1
Std dev	7.1	8.6

the first set (Ford1, Nick1, Laser1, Tax1), using the correctly typed source text. I answered “No” to the prompt that asked whether the analyzed text was representative of the speaker’s style. This caused the system to add all OOV words, but no associated n -gram information, to the user cache.

Results. Table 2 contains the raw data by speaker. For related text, the effect of adding OOV words to the user cache without associated n -gram information was a significant reduction in word error rate of 7.0% ($t(7) = 2.77$, $p = .03$). The percentage of OOV calculated for this text in Table 1 was 0.5%. However, only half of this amount (0.25%) contained words added to the vocabulary when doing vocabulary expansion on the first half, so the remaining portion of the improvement must be due to other factors. The mean percentage reduction in word error rate for unrelated text was a nonsignificant degradation of -4.1% ($t(7) = 1.35$, $p = .22$).

Even without updating LM data in the user cache, vocabulary expansion improved the dictation accuracy of text related, but not identical, to the text processed with the vocabulary expander. The treatment had no significantly adverse effect on the dictation accuracy of unrelated text.

Experiment 2: Normal vocabulary expansion with user cache language model updating.

Motivation. The goal of this experiment was to characterize the effect of letting the vocabulary expansion procedure add OOV words and n -gram information from the analyzed text to the user cache. This had the effect of combining the influence of adding OOV words and all available LM information to the

Table 3. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 2.

Speaker	Related	Control
1	9.5	-35.0
2	20.1	0.0
3	3.0	3.5
4	-0.8	-19.5
5	30.9	-0.4
6	11.5	15.4
7	4.3	-4.6
8	0.7	-3.6
Mean	9.9	-5.5
Std dev	10.9	15.3

user cache—a less conservative approach to vocabulary expansion.

Method. The procedure for this experiment was identical to that of Experiment 1, except I answered “Yes” to the prompt that asked if the analyzed text was representative of the speaker’s style. Doing this added all OOV words to the user cache and updated the user cache with the n -gram information from the analyzed text.

Results. Table 3 contains the raw data by speaker. For related text, the effect of adding OOV words to the user cache with associated n -gram information was a significant reduction in word error rate of 9.9% ($t(7) = 2.58$, $p = .04$). The mean percentage reduction in word error rate for the unrelated text was a nonsignificant degradation of -5.5% ($t(7) = 1.02$, $p = .34$).

Vocabulary expansion with user cache updating improved the dictation accuracy of text related, but not identical, to the text run through the vocabulary expander. The treatment had no significantly adverse effect on the dictation accuracy of unrelated text. For the following experiments, vocabulary expansion included both the addition of new words and updating of the user cache.

Experiment 3: Normal correction.

Motivation. The goal of this experiment was to characterize the effect of normal correction on the future dictation accuracy of related and unrelated texts. This provided an estimate of the benefit gained by the practice of correcting dictated text using the ViaVoice correction dialog.

Table 4. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 3.

Speaker	Related	Control
1	23.1	-22.3
2	8.5	-5.5
3	8.9	2.9
4	4.0	-2.2
5	22.0	-5.7
6	4.9	14.0
7	-0.8	2.6
8	7.9	-2.6
Mean	9.8	-2.3
Std dev	8.4	10.3

Method. The procedure to assess the effect of normal correction was a little different. One of the recordings for the first set of scripts was run through the customized transcription function. After transcribing the file, all misrecognitions were corrected in all four of the treatment files. This resulted in the types of changes to the user cache that happen as a consequence of normal correction—adding OOV words and providing LM updating for the text in the immediate vicinity of the correction (generally plus and minus two words from the target word). The eight test files for that participant were then run through the customized transcription function.

Results. Table 4 contains the raw data by speaker. For related text, the effect of normal correction was a significant reduction in word error rate of 9.8% ($t(7) = 3.28, p = .01$). The mean percentage reduction in word error rate for the unrelated text was a nonsignificant degradation of -2.3% ($t(7) = 0.64, p = .54$).

Thus, normal correction improved the dictation accuracy of text related, but not identical, to text previously dictated and corrected, and with roughly the same magnitude of effect as vocabulary expansion with LM updating of the user cache (the situation studied in Experiment 2). The treatment had no significantly adverse effect on the dictation accuracy of unrelated text.

Experiment 4: Automatic vocabulary expansion of uncorrected dictation for enrolled speakers.

Motivation. Experiments 1 and 2 investigated the effect of performing vocabulary expansion on existing documents, finding and adding all OOV words and either adding (Experiment 2) or failing to add

(Experiment 1) the LM data from the analyzed documents to the user cache. Experiment 3 investigated the effect of normal correction on dictated text. None of these experiments addressed the possibility that a user might save a file of uncorrected dictation with the intention of correcting misrecognized text at a later time. Because the proposed design was to perform automatic document analysis when the user saved a file, there was concern that leaving the uncorrected text in the document might contaminate the LM data in the user cache, leading to reduced accuracy in future dictation. The purpose of Experiment 4 was to characterize the effect of this scenario on future dictation.

Method. The procedure for this experiment was a blend of the procedures used in Experiments 2 and 3. One of the recorded files for the first set of scripts was run through the customized transcription function. After transcribing the file, the uncorrected dictated text from the first half only was copied into a separate file. This was done for all four files, adding the uncorrected first half for each document into the same file. Then, after restoring a participant's enrollment (with no data in the user cache), the file with the uncorrected dictation was run through Vocabulary Expander. This had the effect of contaminating the user cache with the uncorrected dictation, simulating what would happen if automatic vocabulary expansion were performed when a user saved a dictation file without having corrected the misrecognitions in the dictated text.

Results. Table 5 contains the raw data by speaker. For related text, the effect of simulated automatic vocabulary expansion of uncorrected dictation for enrolled

Table 5. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 4.

Speaker	Related	Control
1	-4.7	-35.0
2	9.6	-4.6
3	-2.5	-3.5
4	0.6	1.1
5	18.1	-4.9
6	-1.6	9.7
7	0.7	1.7
8	-15.4	7.5
Mean	0.6	-3.5
Std dev	9.9	13.8

speakers was a nonsignificant reduction in word error rate of 0.6% ($t(7) = 0.17$, $p = .87$). The mean percentage reduction in word error rate for the unrelated text was a nonsignificant degradation of -3.5% ($t(7) = 0.72$, $p = .49$).

Even though contaminated with uncorrected misrecognitions, the simulated automatic vocabulary expansion did not cause any apparent degradation in recognition accuracy for the related or unrelated text.

Experiment 5: Automatic vocabulary expansion of uncorrected dictation for unenrolled speakers.

Motivation. Despite the success of Experiment 4, there was a fear that if recognition accuracy was lower (for example, if users failed to enroll), automatic vocabulary expansion of uncorrected text might cause a serious degradation in future accuracy. The purpose of this experiment was to provide an estimate of the effectiveness of automatic vocabulary expansion for the situation in which speakers have not enrolled and do not correct their errors immediately, thus putting uncorrected dictation into the user cache at a higher rate than that of Experiment 4.

Method. Other than using a speaker-independent (unenrolled) acoustic model, the method of this experiment was the same as that of Experiment 4.

Results. Table 6 contains the raw data by speaker. For related text, the effect of simulated automatic vocabulary expansion of uncorrected dictation for unenrolled speakers was a nonsignificant increase in word error rate of 0.3% ($t(7) = 0.15$, $p = .88$). The mean percentage reduction in word error rate for the unrelated

text was a nonsignificant improvement of 0.2% ($t(7) = 0.13$, $p = .90$).

Even though contaminated with an average of about 15% word errors, the simulated automatic vocabulary expansion did not cause any apparent degradation in recognition accuracy for either related or unrelated text.

Experiment 6: Automatic vocabulary expansion of corrected dictation for unenrolled speakers.

Motivation. The purpose of this experiment was to investigate the possibility that if the accuracy was lower (for example, if users failed to enroll), automatic vocabulary expansion of corrected text might improve accuracy to an even greater extent than occurred in Experiment 2.

Method. Other than using a speaker-independent (unenrolled) acoustic model, the method of this experiment was the same as that of Experiment 2.

Results. Table 7 contains the raw data by speaker. For related text, the effect of simulated automatic vocabulary expansion of corrected dictation for unenrolled speakers was a significant reduction in word error rate of 6.7% ($t(7) = 7.79$, $p = .0001$). The mean percentage reduction in word error rate for the unrelated text was a nonsignificant improvement of 1.9% ($t(7) = 0.81$, $p = .44$).

Accuracy improved for the related text. The treatment had no significantly adverse effect on the dictation accuracy of unrelated text. The data did not support the hypothesis that the amount of improvement would be greater as a function of lower baseline accuracy.

Table 6. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 5.

Speaker	Related	Control
1	-1.22	-5.26
2	1.76	5.35
3	-13.68	-4.43
4	4.52	3.75
5	-2.34	-2.34
6	7.63	-3.08
7	0.54	4.80
8	0.17	2.77
Mean	-0.3	0.2
Std dev	6.3	4.4

Table 7. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 6.

Speaker	Related	Control
1	7.6	-3.2
2	5.1	12.0
3	6.2	-2.8
4	5.7	6.7
5	7.8	-5.1
6	10.3	-2.0
7	2.3	10.2
8	8.6	-0.7
Mean	6.7	1.9
Std dev	2.4	6.7

Experiment 7: Automatic vocabulary expansion for dictation contaminated with linguistijunk.

Motivation. The purpose of this experiment was to address a concern about a scenario that, while unlikely, could happen with automatic vocabulary expansion. The proposed design for ViaVoice’s automatic vocabulary expansion was to analyze documents only when users saved dictation files. In most cases, if a file contained a lot of linguistijunk (quasi-random words produced by the recognizer when it interprets background noise as dictated text), then a user wouldn’t save it. If a user did save such a file, though, then the linguistijunk would contaminate the user cache. The prevailing belief was that, because linguistijunk would follow patterns established in the language model, contaminating the user cache with linguistijunk would not affect recognition accuracy. Despite this belief, it seemed prudent to investigate what would happen.

Method. To investigate the amount of damage such contamination could do to future recognition accuracy, a dictation session was started with the microphone placed next to a speaker playing music. Over the course of about 30 minutes, this produced a file with 1582 words of linguistijunk (see the Appendix for the first 250 words). To get linguistijunk treatment scores, the following actions were done for each speaker:

- Restored the speaker’s enrollment
- Ran the linguistijunk file through vocabulary expander
- Ran the test scripts selected for that speaker through the transcription system
- Scored the results

Note that the concept of related text doesn’t apply to linguistijunk contamination because the linguistijunk has no relationship to any of the test texts. For this evaluation, the comparison was that between the baseline and contaminated word accuracy scores for fifteen test scripts (selected using the criterion of sampling from a wide range of baseline accuracies).

Results. The correlation between the baseline and contaminated accuracy scores was very high ($r = .99$, $p = .0000001$). The mean percentage reduction in word error rate from the baseline to the contaminated cache scores was a nonsignificant 1.6% ($t(14) = 0.88$, $p = .39$).

Table 8. Percentage reduction in word error rate: Raw data and basic statistics for Experiment 7.

Test case	Control
1	-2.4
2	8.3
3	-13.1
4	6.1
5	6.0
6	16.9
7	-4.5
8	9.1
9	0.0
10	0.0
11	3.6
12	-5.9
13	0.0
14	3.7
15	-3.2
Mean	1.6
Std dev	7.3

This experiment addressed the concern about a potentially harmful consequence of automatic vocabulary expansion. Apparently, because the language model plays a significant role in what the recognizer produces as linguistijunk, the introduction of this type of text into the user cache had no effect, either harmful or beneficial, on future dictation accuracy.

Summary of Results

Figure 1 summarizes the results for the seven experiments, shown as thirteen 95% confidence intervals. The significant effects are those that do not include the value of 0 in the confidence interval. Given corrected text (Experiments 1, 2 and 6), vocabulary expansion improved the accuracy of subsequently dictated text related to the treated text with roughly the same magnitude as standard correction (Experiment 3), whether or not users had enrolled. Given uncorrected text (Experiments 4 and 5), vocabulary expansion did no apparent harm, whether or not users had enrolled. Contamination of the user cache with linguistijunk (Experiment 7) neither enhanced nor degraded future recognition accuracy.

The data from Experiments 2, 4, 5 and 6 provide all the information required to run a repeated-measures

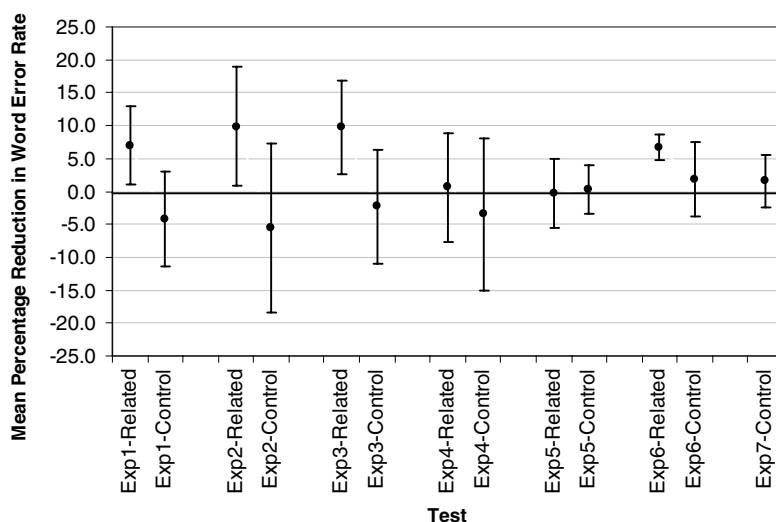


Figure Notes – Critical Features of Experiments:

- Experiment 1: Vocabulary expansion without *n*-gram updating, with enrollment, with correction
- Experiment 2: Standard vocabulary expansion, with enrollment, with correction
- Experiment 3: Standard correction with correction dialog, with enrollment, with correction
- Experiment 4: Standard vocabulary expansion, with enrollment, without correction
- Experiment 5: Standard vocabulary expansion, without enrollment, without correction
- Experiment 6: Standard vocabulary expansion, without enrollment, with correction
- Experiment 7: Standard vocabulary expansion, contamination of cache with linguistjunk

Figure 1. Summary of results.

analysis of variance with three independent variables: Enrollment (enrolled or unenrolled speaker), Correction (misrecognitions corrected or left uncorrected), and Relationship (dictated text related to treatment text or unrelated control text). The analysis indicated a significant main effect for Correction ($F(1,7) = 7.4, p = .03$) and a significant Correction by Relationship interaction ($F(1,7) = 17.1, p = .004$). No other main effects or interactions were significant (all $p > .23$ except for the main effect of Relationship, for which $p = .105$). Figure 2 illustrates the Correction by Relationship interaction. The patterns shown in Figure 1 provide the details necessary to interpret the interaction. Vocabulary expansion improves recognition accuracy, but only if the user has corrected misrecognitions before analyzing the text and the future text bears some relationship to the analyzed text. Analysis of uncorrected text, however, does not result in significant degradation of future recognition accuracy.

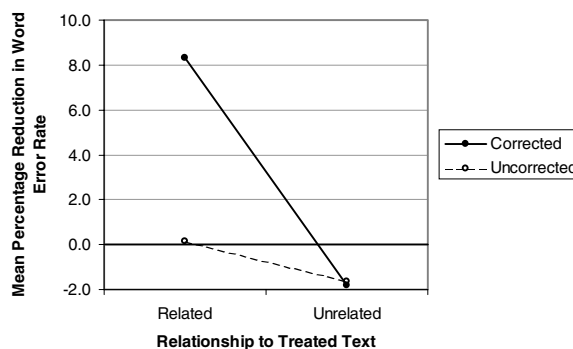


Figure 2. Correction by relationship interaction.

Discussion

An interesting pattern emerged from the results of these seven experiments. Vocabulary expansion on corrected text improved the accuracy of text dictated in the future when the future text was similar to the analyzed

text. None of the manipulations had a measurable effect (either good or bad) when the analyzed text was uncorrected dictation (including linguistijunk). These results were the same regardless of whether the system interpreting the speakers' audio used a trained or untrained acoustic model (in other words, with or without enrollment).

Based on these results, the IBM development team decided to include automatic vocabulary expansion in the ViaVoice product line, starting with ViaVoice Millennium. The major consequence of this to a user of this speech dictation product is that it is no longer necessary to use the correction dialog to receive the long-term benefits of correcting dictation errors. Users can, if they choose, make their corrections directly in the text of the document, using either the keyboard or dictation. This design addresses a major usability problem associated with commercial speech dictation products, as reported by Karat et al. (1999). Furthermore, because there are fewer dialogs to manipulate when editing text directly in a document, it is likely that the average speed of correction would also improve, leading to a substantial improvement in text throughput when using speech dictation to produce a document (Lewis, 1999).

Although the results of Lewis (1999) and Karat et al. (1999) indicate that this change should have the effect of improving user throughput when dictating text and user satisfaction with a speech dictation product, the extent of the improvement, if any, is currently unknown. For this reason, an important item for future work is to replicate the study of Lewis (1999), including a condition in which users make corrections directly in the text rather than using the correction dialog.

Appendix: Linguistijunk Text—First 250 Words

Her when his who hopped all all left when he her half his his and 50 unhinge his inhale inhibit inhibit uphill has been reflecting telescope his inhibitions if hands-off and her left on his physician has a half-life has often leaving only half who is enough whose his his his has behaved him fajita fifth-behaved himself has fajita him him has an Atlanta who often behave after his defeat federal himself inherent final and unless his fists him why is still laugh if it has had half's high-heeled him how how has how has asked has her hand denounce has has has half-has how his has in in in in in in in in in in in this has half has our our our our our our household hour-and-a-half he had him and and and his

hangar who demand has hand-in-hand him how he had his has has only an inland half its inherent when life who flood of whom friends has had her half her how his his half Howe himself has House whip a half-million physician has has he wouldn't soon in Nazi who falls in him his uninhibited flew him him home who and handed Hines and ended and landed in had at an has his is ruined him fluffs fat-free and fluff half an Iraqi who has this half-lift him for now if NASA had him in a handful of his half a hint his hoofs has himself has sanctioned and has handed him

Acknowledgment

Many thanks to Michael Picheny for his help in acquiring the technical information on the recognition methods used in ViaVoice.

References

- Abelson, R. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum.
- Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., and Picheny, M.A. (1988). Acoustic Markov models used in the Tangora speech recognition system. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY: IEEE, pp. 497–500.
- Bahl, L.R., Das, S.K., de Souza, P.V., Epstein, M., Mercer, R.L., Merialdo, B., Nahamoo, D., Picheny, M.A., and Powell, J. (1991). Automatic phonetic baseform determination. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada: IEEE, pp. 173–176.
- Das, S.K. and Picheny, M.A. (1996). Issues in practical large vocabulary isolated word recognition: The IBM Tangora system. In C.H. Lee, F.K. Soong, and K.K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Boston, MA: Kluwer Academic Publishers, pp. 457–479.
- Jelinek, F. (1999). *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press.
- Karat, C.M., Halverson, C., Horn, D., and Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. *CHI '99 Conference Proceedings*, Pittsburgh, PA: Association for Computing Machinery, pp. 568–575.
- Lewis, J.R. (1999). Effect of error correction strategy on speech dictation throughput. *Proceedings of the Human Factors and Ergonomics Society*, Santa Monica, CA: Human Factors and Ergonomics Society, pp. 457–461.
- Lucassen, J.M. and Mercer, R.L. (1984). An information-theoretic approach to the automatic determination of phonetic baseforms. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY: IEEE, pp. 42.5.1–42.5.4.
- Steele, R.G.D. and Torrie, J.H. (1960). *Principles and Procedures of Statistics*. New York, NY: McGraw-Hill.