ACES: Aphasia Emulation, Realism, and the Turing Test

Joshua Hailpern
Department of Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61801
jhailpe2@cs.uiuc.edu

Marina Danilevsky
Department of Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61801
danilev1@illinois.edu

Karrie Karahalios
Department of Computer Science
University of Illinois
201 N Goodwin Ave
Urbana, IL 61801
kkarahal@cs.uiuc.edu

ABSTRACT

To an outsider it may appear as though an individual with aphasia has poor cognitive function. However, the problem resides in the individual's receptive and expressive language, and not in their ability to think. This misperception, paired with a lack of empathy, can have a direct impact on quality of life and medical care. Hailpern's 2011 paper on ACES demonstrated a novel system that enabled users (e.g., caregivers, therapists, family) to experience first hand the communication-distorting effects of aphasia. While their paper illustrated the impact of ACES on empathy, it did not validate the underlying distortion emulation. This paper provides a validation of ACES' distortions through a Turing Test experiment with participants from the Speech and Hearing Science community. It illustrates that text samples generated with ACES distortions are generally not distinguishable from text samples originating from individuals with aphasia. This paper explores ACES distortions through a 'How Human' is it test, in which participants explicitly rate how human- or computer-like distortions appear to be.

Categories and Subject Descriptors

H.5.0 [**General**]; K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities

General Terms

Experimentation, Human Factors

Keywords

Aphasia, Assistive Technology, Disabilities, Empathy, Emulation Software, Language, Speech, Turing Test

1. INTRODUCTION

If a traveler visits a foreign country whose language is not her own, there is no social expectation that she can speak the local tongue. Travelers often get the benefit of the doubt, because their conversation partners have been in a similar situation, and can empathize with the challenges of being in a foreign country. However for more than one million individuals with Aphasia [15], the "foreign language" is their native tongue. Aphasia is an acquired language disorder caused by damage to the left or dominant hemisphere of the brain (often associated with strokes). The disorder impairs an individual's ability to produce and understand language in both written and spoken forms [1]. To an outsider it may appear that an aphasic individual has poor cognitive function. However, the problem resides in the individual's receptive and expressive language, and not in their ability to think. Unfortunately, many friends and family avoid interacting with individuals with aphasia because they do not understand the disorder, lack empathy, and simply find interacting to be difficult. This lack of empathy can "erode the social bonds that give life meaning," and greatly diminish quality of care in a professional setting (e.g. by doctors and nurses) [11].

In 2011, we published ACES (**A**phasia **C**haracteristics **E**mulation **S**oftware), a novel system that enables users to experience the speech-distorting effects of aphasia [9]. ACES uses a probabilistic model (based on literature in Cognitive Psychology and Speech and Hearing Science) that can distort a user's Instant Messages (IMs), transforming the original message into text that appears as though it had originated from an individual with aphasia. Results from an evaluation of 64 participants indicated that ACES strongly increased understanding and empathy for aphasia, and individuals with aphasia.

While our original ACES paper presented the first language disorder emulation system and its impact on empathy, it did not validate the quality or realism of the distortions ACES applied. This paper seeks to demonstrate how discernible ACES distortions are from actual statements generated by individuals with aphasia. If we demonstrate that distortions users experienced were realistic, we will increase the impact and validity of our original study. Further, if ACES appears to be nearly indistinguishable from realistic distortions, it will indicate that ACES may prove a valuable and realistic aid for increasing empathy for family members, friends, clinicians in training and other caregivers.

This paper illustrates the "realism" of ACES distortions in two ways. First, we perform a Turing Test study in which participants must distinguish samples of distorted text generated by a human from samples of text distorted by a computer. Much like the original Turing Test proposed by Alan Turing [23], if participants cannot reliably tell the origin of distortions (whether computer or human generated), the computer could be said to have passed the test. Second,

ASSETS 2011 Dundee, Scotland

we ask participants to explicitly rate the realism of distortions in text samples on a Likert scale. If participants rate both computer and human generated text samples as being equally realistic, it adds further quantitative support to the realism of ACES distortions. The foremost contribution of this paper is the demonstration that ACES generates realistic aphasic distortions, thus validating the applicability of ACES as aphasic emulation software and supporting the feasibility of other language emulation software research.

2. RELATED WORK

We describe aphasia, the theory of linguistic changes in conversation and how our work builds upon, and extends, the existing literature.

2.1 Aphasia

Aphasia is an acquired language disorder that results from by damage to the left or dominant hemisphere of the brain. It impairs an individual's ability to produce and understand language in both written and spoken forms [1]. Because the severity and pattern of aphasic symptoms vary, classification systems were developed to identify different sub-types of aphasia. For example, diagnostic batteries [8, 19] based on the Boston classification system are designed to categorize an individual's aphasia symptoms as either a type of non-fluent aphasia (Broca's, Transcortical Motor, Global) or fluent aphasia (Wernicke, Transcortical Sensory, Conduction, Anomic). Across all sub-types, aphasia impairs the ability to generate written text (though the degree of impairment varies). It should be noted that the linguistic deficits in writing will generally be consistent with those of the person's spoken language [2].

2.2 Empathy and Aphasia

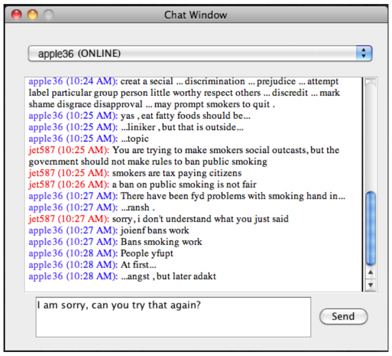
Empathy is one of the fundamental underpinnings of interpersonal communication. It is an emotional response to the experiences of others, through which an empathetic person can understand and predict the feelings, emotions, and thoughts of others [6, 22].

If individuals relating to those with aphasia lack empathy and understanding, it may greatly reduce quality of life for aphasic individuals [11]. Often, family members deny or underestimate the severity and presence of aphasic errors [4]. Further, in speech therapy, empathy is necessary to motivate the aphasic client, with motivation being one of the three key aspects of effective treatment [20]. To date, research has shown that family members' ability to relate and empathize is based on how well they understand the distortions that their family member makes [7].

3. ACES

Motivated by the need to maintain social bonds for those with aphasia, we built a system called **A**phasia **C**haracteristics **E**mulation **S**oftware **(ACES)**, that enabled a user to experience the speech-distorting effects of aphasia first hand [9]. ACES introduced a novel system and model that enabled users (e.g., caregivers, speech therapists and family) to experience, firsthand, the communication-distorting effects of aphasia. The ACES system was designed to distort a user's Instant Messages (IMs) from the original message to one that appeared like a message spoken/written by an individual with aphasia (see Figure 1). Thus, the conversation that developed between the user and their IM partner has similar difficulties and hurdles to those experienced by an individual with aphasia. Similar to spending a day in a wheelchair so as to heighten awareness of the challenges confronting a paraplegic [5], ACES allows a neurologically typical individual to "walk in the shoes" of an individual with aphasia. The goal of ACES was to increase empathy, awareness and understanding of individuals with aphasia. The design and motivation of ACES was grounded in speech and hearing science and psychology literature, and informed by an initial pilot study. Results from a study of 64 participants indicated that ACES provided a rich experience that increased understanding and empathy for aphasia.

While full details of ACES, and the original experiment can be found in the original paper [9], we briefly highlight the key aspects of ACES and our original study that tests ACES' impact on empathy.



```
apple36 (10:24 AM): creat a secial ... discrimination ... prejudice ... attempt
label particular group person little worthy respect others ... discredit ... mark
shame disgrace disapproval ... may prompt smokers to quit .
apple36 (10:25 AM): yas , eat fatty foods should be...
apple36 (10:25 AM): ...liniker , but that is outside...
apple36 (10:25 AM): ...topic
jet587 (10:25 AM): You are trying to make smokers social outcasts, but the
government should not make rules to ban public smoking
jet587 (10:25 AM): smokers are tax paying citizens
jet587 (10:26 AM): a ban on public smoking is not fair
apple36 (10:27 AM): There have been fyd problems with smoking hand in...
apple36 (10:27 AM): ...ransh .
jet587 (10:27 AM): sorry, i don't understand what you just said
apple36 (10:27 AM): joienf bans work
apple36 (10:27 AM): Bans smoking work
apple36 (10:28 AM): People yfupt
apple36 (10:28 AM): At first...
apple36 (10:28 AM): ...angst , but later adakt
```

I am sorry, can you try that again?

Figure 1: ACES Instant Message Window. The current (red) user‘s partner (blue) has their text distorted. This is a log from an actual experiment conversation

3.1 System Design

ACES is a configurable probabilistic model of the linguistic distortions associated with aphasia situated in an IM client. To emulate the distortions generated (and probability of occurance), we leveraged the vast literature in psychology and speech and hearing science (e.g. Schwartz et. al.[18] and Sarno's *Acquired aphasia* [17]). An initial pilot study in 2009 with 10 faculty and practitioners in Psychology and Speech and Hearing Science further informed the system design. Using these as the basis of our models and manifestations of the distortions, we created a system that could emulate a wide variety of distortions across five conceptual categories of distortions. We also allow a user to change the severity and type of aphasia they wished to emulate, thus giving ACES increased flexibility and a broader set of applications.

To a user, ACES appears as an IM client. However, a user's original text is changed to appear like a message written/spoken by an individual with aphasia (see Figure 1). Thus, the conversation that develops between the user and their IM partner has similar challenges to those experienced by an individual with aphasia. In effect, users can "walk in the shoes" of an individual with aphasia. We envisioned ACES to have a multitude of applications; from teaching empathy and awareness to therapists, to helping family and friends understand what their loved ones are experiencing.

ACES was implemented using Java 1.6, allowing the software to execute on nearly any machine. The IM protocol was facilitated by the JBuddy Library [21].

3.2 Empathy User Studies

Upon completion of the ACES system in 2010, a detailed study was conducted. The focus of this in-depth user study was to observe the effects of ACES on awareness and empathy. Sixty-four participants were recruited (both experts on aphasia and individuals from the general population), and tasked to have two IM conversations using ACES in which one participant played the "role" of an individual with aphasia (having their text distorted by ACES so that it appeared generated by an individual with aphasia) while the other participant played the role of a "neurologically typical" individual (non-distorted text). Participants switched role after the first conversation. One half of the participants were assigned to the placebo or **control group** (no distortions applied) and one half were in a **treatment group** (ACES distorted messages sent), in a between subject design.

Participants in the treatment group, who experienced the distortions of aphasia, had a stronger response than those in the control group. Across the metrics used, there was a highly significant effect of ACES on empathy, awareness and how "useful" ACES would be for building empathy. While participants in the treatment group reported "strong" effects from the experiment, those in the control group reported little or no change in their empathy towards aphasia[1][9].

4. RESEARCH QUESTION & MOTIVATION

The initial ACES user study examined the impact of ACES on empathy and awareness of aphasia. However, there was no examination of the distortions themselves, their perceived quality, or their realism. While ACES' distortions were deeply grounded in literature (providing both the probabilistic underpinnings and the manner in which text is distorted), creating a novel and unique form of language emulation has the potential to produce distortions at varying degrees of realism. Therefore this paper seeks to answer the following questions:

> **Can users differentiate computer-generated distortions from distortions generated by individuals with aphasia?**
>
> **How realistic are the distortions of aphasia generated by ACES?**

We answer these questions with a two-step experimental design, detailed in the following section. We then describe our target population, outline the methods for analysis, and present results for each of our two experiments. Discussion of our results follow, along with implications for future work.

5. EXPERIMENTAL DESIGN

To answer our research questions, we recruited 24 participants to examine distortions generated by ACES, utilizing an online questionnaire. Each questionnaire presented users with a set of demographic questions, in addition to 48 data generating questions. Each page of the online questionnaire contained only one data generating question. A data generating question consisted of a text sample and a question about the sample. One half of the text samples were generated by ACES, while the other half were taken from transcripts generated by individuals with aphasia. These two halves will be referred to as the **Computer Group** and the **Human Group**, respectively. In an experimental context, these can be thought of as a treatment group and a control group. If ACES distortions are indistinguishable from human generated distortions, which is the goal of this project, *the desired outcome is to see a lack of statical significance ($p \geq 0.05$)* when comparing the Computer Group to the Human Group.

To control for order effects, we presented the questions in one of two sequences. The question order within each sequence was created randomly. Users were randomly assigned to one of the two presentation sequences. No text sample was presented to a user more than once (to control for learning effects). At the end of the study, participants received a $7 Amazon.com gift certificate.

The next section describes the types of aphasia targeted in this study. We then detail the two sets of questions asked of participants, and discuss the origin of all text samples presented to users.

5.1 Types of Aphasia

Based on the Boston classification system, there are numerous types of aphasia that can be broadly categorized as non-fluent aphasia (Agrammatic, Broca's, Transcortical Motor, Global) or fluent aphasia (Wernicke, Transcortical Sensory, Conduction, Anomic). Individuals within each subtype of aphasia have distinctive characteristics to their speech, and the errors that are made. We therefore wished to control for aphasia type in this experiment. Rather than tackling all known subtypes, we focused on two of the more common **Types of Aphasia**: Agrammatic and Anomic aphasia. Individuals with Agrammatic aphasia generally have difficulty with sentence structure and proper grammar, while having no difficulty with word selection. Common errors include difficulties in verb tense, dropping function words, inconsistency with the length or fluidity of sentences, and incorporating many breaks and pauses. Individuals with Anomic aphasia have difficulty with selecting and producing correct content words, though their grammar is generally correct. For example, words may be replaced by other words that are semantically related ('birthday' with 'anniversary' or 'cake'), that have no semantic relationship (cat with airplane), that have similar phonetic sounds ('population' with 'pollution'), or non-words ('castle' with 'kaksel'). The availability of Anomic and Agrammatic aphasia transcripts, in addition to their general prevalence of individuals with these types of aphasia, influenced our selection.

5.2 Test Questions

The 48 data generating questions were evenly divided into two distinct **Tests**: *The Aphasia Turing Test* and the *"How Human" Test.* Each user first answered the 24 *Turing Test* questions, followed by 24 *"How Human"* questions. Each question was presented on a separate page. This limited participants ability to make judgments based on the other questions' distortions. Further, participants were unable to return to prior questions, thus preventing them from chang-

[1] There were no statistical differences in the effect of ACES between experts in aphasia and individuals from the general population.

ing their answers. The questions and the presentation of each Test are detailed in the following sub-sections. At the end of each test, we asked users to describe their approach for answering the Tests' questions.

5.2.1 Aphasia Turing Test

For the *Aphasia Turing Test*, participants were presented with 24 text samples. For each text sample participants were instructed, *"for each sentence, please mark if it is 'Human' or 'Computer' in origin."* The following example is an actual text sample used in this test, with the distortion generated by ACES emulating an anomic individual with aphasia:

> *Well she was a kaur girl and she qobred in a house where she was mopping the under foot. Then there was something about a shoe and when she wore it she would be Cinderella. uh... She was told that a mumpkur would be her stagecoach... um... and little rats would be a horse. And so she went up to the castle. Her new shoes um... uh... fit uh... um...her like a... grove.*

Participants *were* told that some text was generated by an actual individual with aphasia (Human Group text samples), and some text was distorted by ACES (Computer Group text samples). Participants were not told it was a 50/50 split of text samples from the Human Group and Computer Group. Section 5.3 details how text samples were generated/collected.

In this regard, this experiment was designed as a variation of the Turing Test as proposed by Alan Turing in 1950 [23]. The goal of this Test was to determine if our subjects could reliably differentiate machine from human. To "pass" the Turing Test, we would expect to see approximately a 50% accuracy at labeling text as computer or human (with no statistical difference in the accuracy between groups). Since modern computers cannot reliably pass the Turing Test, we did not hypothesize a priori that ACES would completely pass our Turing Test either. Even without passing the Turing Test, results can illuminate the believability of ACES' distortions, and if there is one type of aphasia (see Section 5.1) which ACES emulates more successfully.

5.2.2 "How Human" Test

For the *"How Human" Test*, participants were presented with 24 pairs of text samples (one pair per page). For each pair of text samples, the first was labeled as "Original Text" and the second as "Distorted Text." The "Original Text" was undistorted, while the "Distorted Text" had aphasic distortions applied to it. The following example is an actual text sample used in this test, with the distortion generated by ACES emulating an agrammatic individual with aphasia:

> **Original Text:** *Well the man is trying to wake up because of the alarm clock. And then he goes back to sleep. His wife is angry. Then the man eats breakfast, while his wife is showing him the time on the clock; the wife is saying "hurry up." And the man running out onto the street to get to work. The man was so tired, he goes to sleep at the office.*
>
> **Distorted Text:** *uh... the er... uh... is tri... wake up because the alarm clock uh... And... he goes back uh... His wife is angry, Then the man eat uh... breakfast, his wife is ah... eh... showing him the time... the wife is saying "um... up." uh... er... And the man running ah... onto the street to get to work. The uh... was so tired he goes to sleep at the office.*

Participants were asked to help researchers "improve" the distortion algorithms by rating how "human" the distortions appear on a Likert scale from 1-5 (where 1 is indistinguishable from a human who has aphasia, and 5 is unquestionably a computer). Like the *Aphasia Turing Test*, one half of the "Distorted Text" samples were generated by an individual with aphasia (Human Group text samples), and the other half of the text samples were distorted by ACES (Computer Group text samples). Section 5.3 details how both the original text and distorted text were generated/collected.

By design, this task forces users to make an implicit judgment call about the origin of each distorted text sample: "was this text distortion generated by a human or by a computer?" To allow participants to focus on the realism of the distortions/errors in the text samples rather than puzzling over their source, participants were told that all text was distorted by ACES. This deception allows us to objectively measure the realism of ACES distortions (Computer Group). It also provides an objective and comparable benchmark of human distortions (Human Group).

5.3 Text Samples

All text samples used in this experiment were extracted from published transcripts of individuals with aphasia [3, 13] or from the unpublished data files used in [12, 14], which were provided to the researchers by Lise Menn, University of Colorado. Some transcripts were from picture describing tasks from the Wechsler Bellevue Intelligence Scale [24]. Other text samples were from transcripts of individuals with aphasia reading children's stories[2]. The remainder of the text samples were from individuals with aphasia narrating children's stories from memory.

For each text sample, an original or intended version of the text was generated. If the transcript was from an individual reading a story, the read text was used as the original version. For transcripts that were not of an individual reading text, researchers attempted to fix the errors and create a non-distorted version of the same text (following similar sentence structure, word choice, and phrasing). For the *"How Human" Test*, these original versions of the text samples were used. The original versions of each text sample were also used as the basis for the ACES distortions. Each non-distorted, or error-free text sample would be sent through ACES, thus applying the ACES distortions to the text. Section 5.3.1 details the procedure for choosing and applying ACES distortions so as to ensure the distorted texts used were not "cherry picked." There were therefore three versions of each text sample, the aphasic version, the 'original' version, and the ACES version. Selection of text samples to be included in the experiment was random, thus not giving any preference to 'more believable' ACES text.

Our text samples came from six individuals. We ensured that there were an equal number of text samples from each

[2] Reading does produce errors in speech production [13].

individual within both Tests (e.g. aphasiac individual Alice contributed four text samples to the Turing Test, and four text samples to the "How Human" Test). Further, the text samples taken from each participant were split evenly across the Human Group and Computer Group (e.g. if aphasiac individual Bob contributed eight text samples, four were used directly from his transcript and four were used to construct an undistorted text sample, which was then distorted by ACES). No text sample was repeated across Tests or within a Test, and only one version (the true aphasic version or the computer version) of each text sample was used.

5.3.1 Generating ACES Distortions

For each set of transcripts generated by one individual, researchers constructed an ACES model that attempted to emulate his or her manifestation of aphasia. This was done by taking text [3], running it through ACES, and adjusting the software's distortion parameters until the distorted text appeared 'similar' to the transcripts that were be generated by said individual. Only the sliders on the ACES interface were adjusted (no code was edited).

Once a model was set, every 'original' version of text sample generated by that individual was then run through ACES once. No text sample was repeated. This ensured that our study used whatever distortions ACES applied, without preference to more 'successful' distortions. These distorted sentences were then cleaned, fixing spacing or punctuation issues that may be a byproduct of removing or adding words. No word spellings, phrasing or other changes were made to the ACES text, further ensuring that the distortions shown to participants were precisely the ones generated by ACES.

5.4 Population

We recruited 24 participants (3 male, 21 female) for inclusion in this study. Participants were students or faculty in Speech and Hearing Science Departments, as well as professionals in the Speech and Hearing Science community. We chose Speech and Hearing Science students, faculty and professionals as our target population because their training is specifically targeted towards the identification and treatment of speech disorders. Part of this training includes analyzing transcripts of conversations, diagnosing disorders based on language production (thereby distinguishing one from another), and treating the speech disorders themselves. We felt that this population was uniquely qualified to perform the discrimination tasks in this experiment.

We actively recruited from multiple institutions to cultivate a wide perspective on aphasia. Of our participants, all had taken at least one class that covered aphasia, and 67% of participants had personal experience with aphasia, or had taken a class that only covered aphasia. The population contained four current BS/BA students, 13 current MS/MA students, five participants with an MS/MA degree, and two participants with a PhD. The mean age of our participants was 26.4 (range 19 to 60 years).

5.5 Analytical Methods

To examine the quality of the ACES distortions, we compared the participants' responses to the 24 *Aphasia Turing Test* questions separately from the 24 responses to the *"How Human" Test* questions. For each Test, we treated all responses to that Test's questions as one uniform data set. Since a participant contributes more than one data point within a test, the responses are correlated. Therefore, statistical tests must take into account the correlated nature of the data. For statistical comparison, we compared responses to text samples in the Human Group with responses to text samples in the Computer Group.This compares participants accuracy in distinguishing human distortions from distortions generated by ACES.

	Text Sample	**Participants' Label**	
	Group	Correctly	Incorrectly
	Human	146 (50.69)	142 (49.31)
Overall	Computer	155 (53.82)	133 (46.18)
	Total	301 (52.26)	275 (47.74)
	Human	87 (60.42)	57 (39.58)
Anomic	Computer	78 (54.17)	66 (43.83)
	Total	165 (57.29)	123 (42.71)
	Human	59 (40.93)	85 (59.03)
Agrammatic	Computer	77 (53.47)	67 (46.53)
	Total	136 (47.22)	152 (52.78)

Table 1: Aphasia Turing Test Results
Occurrence count with row percentages (accuracy) in parentheses

It is important to note that in this experiment, lack of statistical significance is the desired outcome. Statistically significant tests results generally, by definition, look for differences. If ACES distortions are indistinguishable from human generated distortions, we would see a lack of statistical significance ($p \geq 0.05$) between the Computer Group data set and the Human Group data set.

Responses to the *Aphasia Turing Test* were binary (users marked each text sample as Human or Computer in origin). This would suggest using a Pearson's Chi-Squared, Fisher Exact or Binomial test. However, these tests do not account for the correlated nature of the data (each participant answered multiple questions that were analyzed collectively). Generalized estimating equations (GEE) [10] with a logistic regression[4] were used to account for these correlations. To augment our analysis, we also examined the percentage of data points that were labeled correctly, and the percentage labeled incorrectly. Lastly, we separated out the Anomic and Agrammatic text samples to determine if aphasia type impacted participants' ability to discriminate.

Responses to the *"How Human" Test* were categorical. This would suggest using a Two-Sample Wilcoxon Rank-Sum (Mann-Whitney) test, a more conservative metric than the Student's T-Test as it makes no assumptions about the data distribution. However, Rank-Sum tests do not account for the correlated nature of the data (each participant answered multiple questions that were analyzed collectively). Generalized estimating equations (GEE) [10] with a linear regression[5] were used to account for correlation.

To further inform our analysis of the *"How Human" Test*, we also examined the distribution of data points. As an explicit measure of similarity of our two data sets, we utilized

[3] To remove bias, text used to calibrate distortions was unrelated to the aphasic transcripts used in this study.

[4] Logistic regressions were used to test associations with binary outcomes (correct/incorrect labeling by participants).

[5] Linear regressions were used to test associations with scale responses (Likert scale 1-5) as outcomes.

Rita and Ekholm's measure of similarity[6][16]. This similarity metric utilizes a θ, or tolerance in the means between two data sets. We set a conservative θ to be one fifth of a Likert interval (0.2). This represents 5% of the possible answer range, and just over one eighth (13%) of the overall variance (1.50) in subject responses to the "How Human" Test Likert questions.

6. RESULTS

Our data set consisted of 1152 observations (data points), from 24 participants. Of these, 576 observations were from the *Aphasia Turing Test*, and 576 were from the *"How Human" Test*. The following sections detail the quantitative results from our analysis.

6.1 Results for Aphasia Turing Test

As shown in Table 1, overall participants correctly discriminated Human vs. Computer slightly better than chance (52.26%). Similarly, within the two text sample Groups, participants correctly categorized slightly over 50% of the text samples. GEE tests indicated no statistical difference between subjects' ability to discriminate text samples from the Human group with text samples from the Computer Group (z= -0.75, p=0.46).

Further analysis of Anomic text samples (Table 1) shows similar results to that of the overall dataset. A comparison of the accuracy of rating the Human Group versus the Computer Group indicated no statistical difference between the two groups (z=1.06, p=0.29). Analysis of the Agrammatic text samples (Table 1) produced different results. Specifically, participant performance dropped considerably in their ability to correctly label text samples from the Human Group: 60% with Anomic text samples, 40% with the Agrammatic text samples (z= -2.08, p=0.04).

We performed a post hoc analysis (GEE test), comparing participants' performance between Anomic distortions and Agrammatic distortions within each text sample group (e.g., Anomic Human Group vs. Agrammatic Human Group) to determine if the participants could differentiate Anomic or Agrammatic better. Results showed no statistical difference between Anomic and Agramatic text samples from the Computer Group (z=0.12, p=0.91). However a highly significant difference was seen between Anomic text samples from the Human Group and Agrammatic text samples from the Human Group (z=3.28, p=0.001). This may indicate that the ability of participants to differentiate Agrammatic text samples that were from the Human Group (41%) was significantly poorer than when examining Anomic text samples from the Human Group (60%).

6.2 Results for the "How Human" Test

Table 2 shows summary statistics and sparklines for the distribution of participant responses to the "How Human" test, ranging from 1 (Definitely Human) to 5 (Definitely Computer). Overall, participants rated ACES generated distortions as 3.05, showing a slight favor towards being computer in origin. Likewise, participants rated text samples from the Human Group as 2.94 overall, showing a slight favor towards being human in origin. However these slight shifts in preference showed no statistical significance (z=-1.17, p=0.24). Using the Rita and Ekholm's similarity measure [16], the two data sets were found to be statistically similar ($p<0.05$).

When we stratify our data by Aphasia Type, our results diverge. For text samples that had Anomic distortions, we observed a statistically significant difference (z=-4.10, $p<0.001$). Participants rated text samples from the Human Group as being more human (2.73) than text samples from the Computer Group (3.26). While these differences are about 1/4 of a Likert point away from a neutral score of 3.0, this result indicates that Anomic distortions were slightly less believable. This is confirmed with Rita and Ekholms' similarity measure ($p \geq 0.05$).

Results of the Agrammatic text samples, however, ran contrary to ground truth. While there was a statistically significant difference between the Human Group and Computer Group (z=2.32, p=0.02), the mean responses were opposite to the origin of the text. Participants rated text samples from the Human Group as being more computer (3.15) than text samples from the Computer Group which were rated more human (2.84). These responses were biased in the wrong direction. It is also true that these results were statistically not-similar using Rita and Ekholms' similarity measure ($p \geq 0.05$).

7. DISCUSSION

In general, our results indicate that ACES provides a realistic set of distortions of aphasia. Our participants overall had difficulty differentiating between the origins of our text samples, and generally rated distortions as being right between definitely computer in origin, and definitely human in origin. Unlike most experimental setups, lack of significance is a positive outcome, validating the realism of ACES distortions. The remainder of this section discusses the specific results from each Test.

7.1 Aphasia Turing Test

Participants were unable to discriminate between distortions generated by humans with aphasia and distortions generated by ACES. In this regard, ACES distortions passed our variation of the Turing Test. With overall accuracies for both the Human and Computer Group hovering around 50% (nearly equivalent to random chance), and no statistical significance found between the two groups, we can conclude that ACES distortions are indistinguishable from those generated by humans with aphasia.

While the accuracy for identifying Anomic text samples from the Human Group rose slightly, the ability to correctly label Anomic text samples from the Computer group remained constant, and we saw no statistically significant difference between the Control and Human Group.

However, the results from Agrammatic text samples demonstrate an inability of participants to correctly identify text samples that originate from humans (41% accuracy). This probability is worse than chance, and is a statistically significant drop-off as compared to the accuracy for Anomic text samples. Further, the ability to correctly identify Agrammatic text samples from the Computer Group remained constant when compared to Anomic text samples(not sta-

[6] Rita and Ekholm measure uses a similarity limit, θ such that the difference in the averages of the two data sets is smaller than θ in absolute value. This can be determined by examining the 95% confidence interval for the difference between the two data sets. If the θ is greater than the right 95% confidence interval *and* -θ is less than the left 95% confidence interval, 'there is a difference' with $p<0.05$.

	Text Sample Group	Mean (St. Dev.)	95% Conf. Int.	Histogram
	Human	2.94 (1.22)	[2.80, 3.08]	
Overall	Computer	3.05 (1.22)	[2.91, 3.19]	
	Total	3.00 (1.22)	[2.90, 3.10]	
	Human	2.73 (1.16)	[2.54, 2.92]	
Anomic	Computer	3.26 (1.21)	[3.06, 3.46]	
	Total	3.00 (1.21)	[2.86, 3.14]	
	Human	3.15 (1.25)	[2.95, 3.36]	
Agrammatic	Computer	2.84 (1.20)	[2.64, 3.04]	
	Total	3.00 (1.24)	[2.85, 3.14]	

Table 2: Summary Statistics and Histogram Sparkline for “How Human” Test
Response Range from 1(Distortions Definitely Human in Origin) to 5 (Distortions Definitely Computer in Origin).
Histogram shows frequency of each Likert scale rating with 1 on the left, and 5 on the right.

tistically significant). Therefore we attribute the only statistically significant difference in the Aphasia Turing Test to participants’ inability to correctly identify text samples from the Human Group, rather than an increase in their ability to identify text from the Computer Group. Taking this into consideration, we continue to see that participants had approximately a 50/50 chance of correctly labeling text samples from the Computer Group, still indicating that participants were unable to distinguish text samples from the Human and Computer Groups.

We can therefore conclude that, across the board, participants are generally unable to distinguish human distortions from ACES distortions, thus passing our variation on the Turing Test. This adds support to the claim that ACES creates realistic distortions of aphasia.

7.2 “How Human" Test

The overall results from the “How Human” Test paralleled those of the Aphasia Turing Test. Participants’ ratings between Human and Computer Group showed no statistical difference. However, differences emerge when data is stratified by Aphasia Type. In general, Anomic distortions in the Computer Group tend to be labeled as more computer-like, while actual distortions in the Human group are correctly marked as being more human. Analysis confirms that this is a statically significant difference.

However, analysis of the text samples with Agrammatic distortions showed that participants generally believed that the ACES distortions were more human (2.84 on Likert scale 1-5), and the real text samples were more likely to come from a computer (3.15 on Likert scale 1-5). This difference was statistically significant. We therefore speculate on the possible causes of this surprising finding. First, our participants may have had difficulty in identifying Agrammatic aphasia. Second, transcripts (ours, or in general) may not have fully captured the nuances of Agrammatic aphasia. Third, the models and distortions ACES used were based on the same literature that is used to teach speech and hearing science students. It is possible that the literature does not fully describe the nature of Agrammatic aphasia. Therefore ACES may more closely match our participants’ expectations of Agrammatic aphasia as compared to actual transcripts.

It is worth noting that mean scores (across Aphasia Type and Text Sample Group) are relatively close to the center of the 5 point Likert scale (equally human and computer). Examination of the distributions (last column of Table 2), reveals a single or double hump bell curve around a value of 3 on the scale. Thus indicating that participants generally were unable to categorize a text samples’ errors as ‘definitely’ human or computer in origin.

Upon further examination, we determined that no one user performed notably better or worse when answering the “How Human” Test, suggesting that the results were consistent across participants. We also examined participants’ qualitative responses, at the end of the “How Human” Test, commenting on how they made their decisions. Surprisingly, participant responses were not consistent. One participant mentioned placement of pauses in sentences, whereas another participant relied upon how ‘obvious’ a semantic replacement was. However, no two participants mentioned the exact same aspect of speech as being a key informative factor. Moreover, many participants’ responses contained a phrase similar to that of participant 23, *“I was really surprised by how realistic the distortions were to me.”*

7.3 Future Work & Limitations

This work represents an important step forward in validating ACES, and it’s impact. As there are many distinctive subtypes of aphasia, the Aphasia Turing Test should be repeated with each of them, to explore the ability of ACES to emulate each specific type of aphasia. This vein of research would also help guide future development of ACES distortions, and improve the quality of the requisite distortions.

In addition, results from the “How Human” Test highlight that participants find Anomic distortions generated by ACES to be slightly more computer than human. However the specific reasons are unclear given the variety of user responses to the general question “How did you make your decisions?” We therefore propose a future investigation into ACES distortions, focusing only on Anomic errors. This study would ask participants to justify their decision on each question, rather than prompt for one reflective statement at the end of the study. This may provide specific insight into why ACES distortions fail and/or succeed.

Lastly, given the surprising Agrammatic text sample results in the “How Human” Test, future investigations need to be conducted as to why ACES distortions appear more human, and real transcripts appear more computer-like. In addition, this test should be repeated to ensure that this result was not in error.

8. CONCLUSION

Empathy and understanding from family members, friends, professionals and caregivers directly impacts the quality of life and quality of care of individuals with aphasia. To this end, Hailpern et. al. developed ACES, a system which allows users (e.g., caregivers, speech therapists and family) to "walk in the shoes" of an individual with aphasia by experiencing linguistic distortions firsthand. ACES' distortion model was directly based on the literature in the fields of Cognitive Psychology and Speech and Hearing Science. While results from an initial experiment illustrate that ACES increases empathy and understanding of aphasia, the original paper did not explicitly validate the distortion model. Our work has made several contributions to address this limitation.

First, this paper shows that participants from the Speech and Hearing Science community, whose training is specifically targeted towards the identification and treatment of speech disorders, cannot consistently differentiate computer and human generated distortions. Second, from our investigation of the realism of ACES distortions, we discover that overall, both human and computer generated distortions appear equally "realistic." However, when stratified by type of aphasia, we can see that ACES' emulation of Anomic aphasia is slightly less realistic than ACES' emulation of Agrammatic aphasia. Third, by validating the distortions used in Hailpern's original experiment, this paper strengthens the original paper's findings, showing that the distortions experienced were believable approximations of aphasia. Lastly, by coupling the results of this paper and those of the original study, we add support to the feasibility of other language emulation research targeting other language deficits.

9. ACKNOWLEDGMENTS

We would like to thank our reviewers, participants, family and friends. We would like to acknowledge the help and support of Yonatan Bisk and Wai-Tat Fu. We would like to thank Julie Hengst for her time and discussions at the beginning of the original ACES project. We would like to thank Gary Dell for his many hours of assistance on all of the ACES research and publications. Lastly, a special thanks to Lise Menn who provided many text samples which were critical to this work being completed.

10. REFERENCES

[1] D. Benson. *Aphasia, Alexia and Agraphia: Clinical Neurology and Neurosurgery Monographs*. Churchill Livingstone, New York, 1979.

[2] D. Benson. *The neurology of thinking*. Oxford University Press, USA, 1994.

[3] F. Boller and J. Grafman. *Handbook of Neuropsychology, 2nd Edition : Language and Aphasia*. Elsevier Science Health Science div, 2001.

[4] P. Czvik. Assessment of family attitudes toward aphasic patients with severe auditory processing disorders. *Clinical Aphasiology Conference*, 1977.

[5] S. Doerksen. Recreation for persons with disabilities (rpm 277). *Pennsylvania State University Department of Recreation, Park and Tourism Management*, Fall, 2009.

[6] R. Dymond. A scale for the measurement of empathic ability. *Journal of Consulting Psychology*, 13(2):127–133, 1949.

[7] E. A. Furbacher and R. T. Wertz. Simulation of aphasia by wives of aphasic patients. *Clinical Aphasiology*, page 227, 1983.

[8] H. Goodglass, Goodglass, and Kaplan. *Boston Diagnostic Aphasia Examination: Stimulus Cards–Short Form*. Lippincott Williams & Wilkins, 2001.

[9] J. Hailpern, M. Danilevsky, A. Harris, K. Karahalios, G. Dell, and J. Hengst. Aces: Promoting empathy towards aphasia through language distortion emulation software. In *Proceedings of the ACM's SIG CHI Conference 2011 Conference.*, CHI 2011, Vancouver, BC Canada, 2011. ACM.

[10] J. Hardin and J. Hilbe. *Generalized estimating equations*. Chapman and Hall/CRC, New York, 2003.

[11] J. Liechty and J. Heinzekehr. Caring for those without words: A perspective on aphasia. *The Journal of Neuroscience Nursing*, 39(5):316, 2007.

[12] L. Menn, A. Kamio, M. Hayashi, I. Fujita, S. Sasanuma, and L. Boles. The role of empathy in sentence production: A functional analysis of aphasic and normal elicited narratives in Japanese and English. *Function and Structure*, pages 317–356, 1998.

[13] L. Menn and L. Obler. *Agrammatic aphasia: A cross-language narrative sourcebook*. John Benjamins, 1990.

[14] L. Menn, K. Reilly, M. Hayashi, A. Kamio, I. Fujita, and S. Sasanuma. The interaction of preserved pragmatics and impaired syntax in Japanese and English aphasic speech. *Brain and language*, 61(2):183–225, 1998.

[15] National Institute on Deafness and Other Communication Disorders. Aphasia. http://www.nidcd.nih.gov/health/voice/aphasia.htm, 2010.

[16] H. Rita and P. Ekholm. Showing similarity of results given by two methods: A commentary. *Environmental Pollution*, 145(2):383–386, 2007.

[17] M. T. Sarno. *Acquired aphasia (Third Edition)*. Academic Press, San Diego, CA, 1998.

[18] M. Schwartz, G. Dell, N. Martin, S. Gahl, and P. Sobel. A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, 54(2):228–264, 2006.

[19] C. Shewan and A. Kertesz. Reliability and validity characteristics of the western aphasia battery (wab). *Journal of Speech and Hearing Disorders*, 45(3):308, 1980.

[20] M. Shill. Motivational factors in aphasia therapy: Research suggestions. *Journal of Communication Disorders*, 12(6):503–517, 1979.

[21] Z. Software. Jbuddy messenger.

[22] E. Stotland. Exploratory investigations of empathy. *Advances in experimental social psychology*, 4:271–314, 1969.

[23] A. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[24] D. Wechsler. Manual for the Wechsler Adult Intelligence Scale. 1955.