

Corpus Studies in Word Prediction

Keith Trnka and Kathleen F. McCoy
University of Delaware

Outline

- background
- the problem
- training corpora
- testing methods
- example application of testing methods

AAC Background

- [Augmentative and Alternative Communication (AAC)
- [communicating with speech and/or motor impairments
- [AAC devices
 - high-tech devices – word/letter/phrase/icon input, speech synthesis output
- [the communication rate divide and fatigue

Example AAC Device

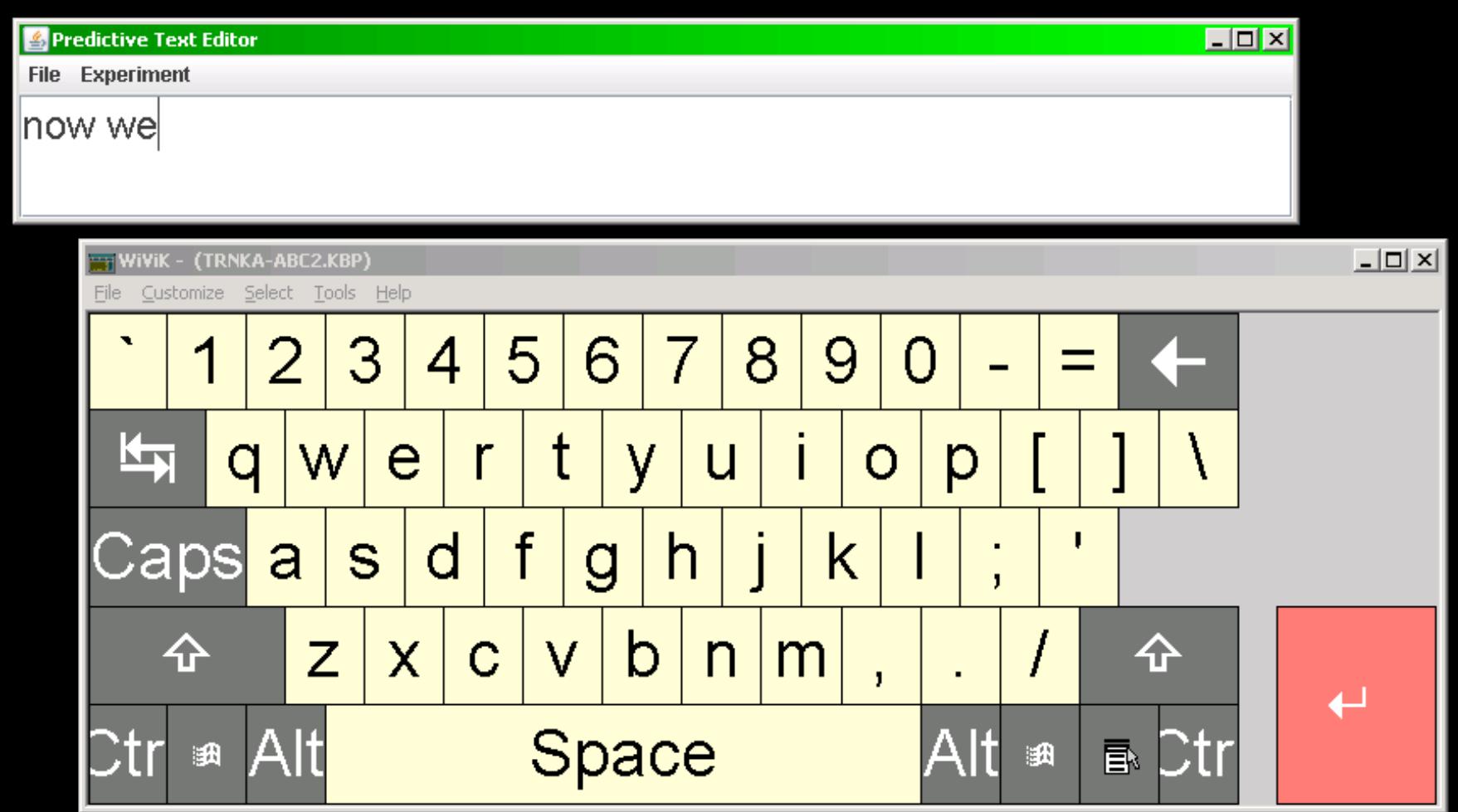
Would you mind if I got a ride home from Jerry after school? I don't want to ride the bus. I am

The image shows a screenshot of an AAC device. At the top, there is a message box containing the text: "Would you mind if I got a ride home from Jerry after school? I don't want to ride the bus. I am". Below the message box is a grid-based keyboard. The keyboard has several rows of buttons. The first row contains buttons for numbers (123, ., !, ?) and punctuation (., !, ?, -). The second row contains letters A through Z, with some letters like Q, E, R, T, Y, U, I, O, P, and L having alternative meanings like 'able', 'hello', 'please', 'thank you', 'shift', 'space', and 'clear'. The third row contains common words like 'be', 'am', 'is', 'are', 'to', 'be', 'call', 'any', 'every', 'about', 'and', 'at', 'THINGS', 'be', 'him', 'can', 'could', 'come', 'get', 'big', 'find', 'some', 'become', 'but', 'by', 'good', 'she', 'her', 'did', 'do', 'get', 'give', 'go', 'help', 'a', 'down', 'for', 'from', 'more', 'they', 'them', 'had', 'has', 'know', 'like', 'make', 'all', 'here', 'if', 'in', 'much', 'you', 'your', 'have', 'is', 'need', 'put', 'say', 'take', 'that', 'all', 'all', 'on', 'really', 'no', 'don't', 'yes', 'were', 'talk', 'tell', 'think', 'use', 'the', 'at', 'out', 'over', 'up', 'Phone 2', 'not', 'will', 'would', 'walk', 'want', 'watch', 'work', 'this', 'there', 'up', 'with', 'very'.

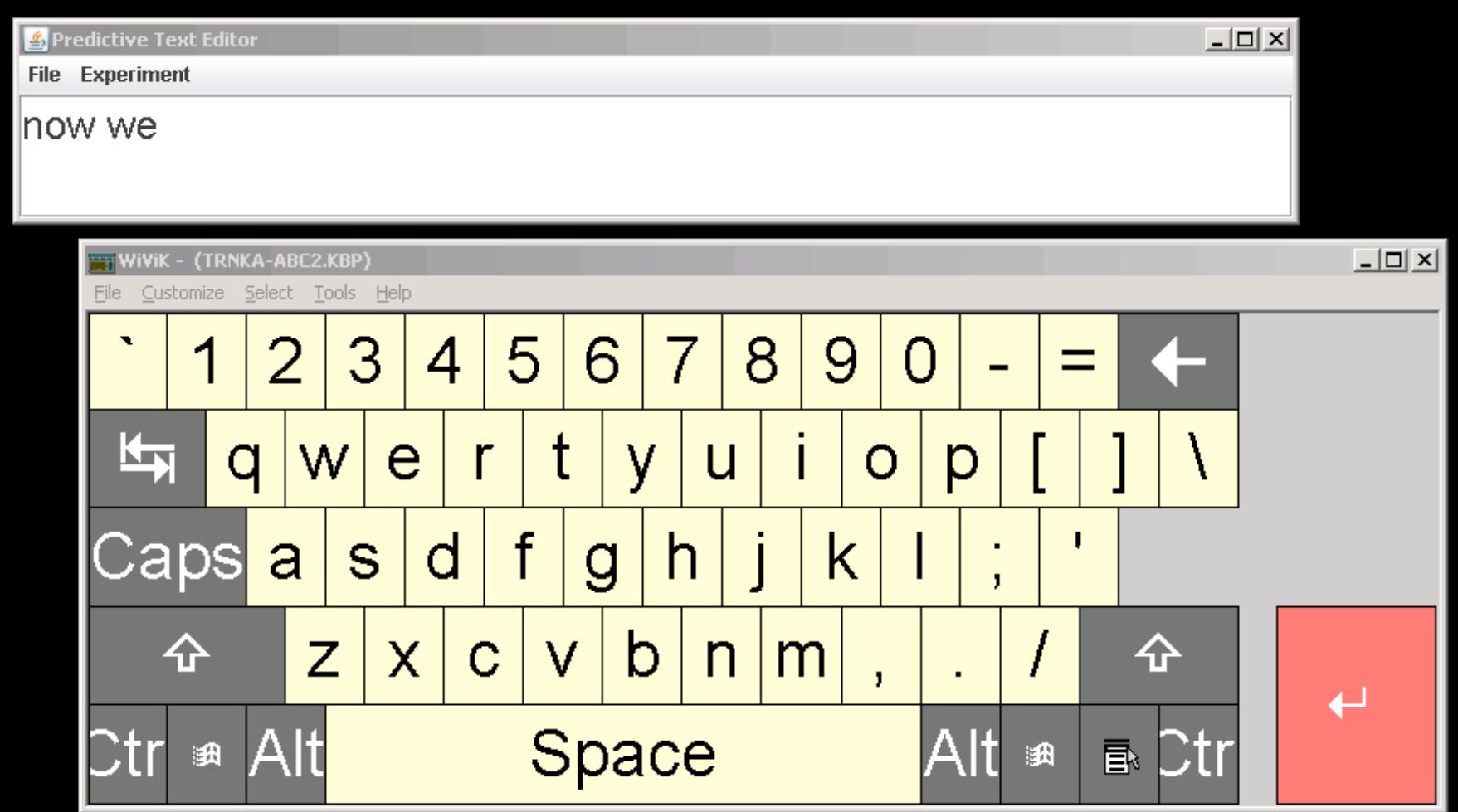
Word Prediction in AAC

- [NLP technique to reduce the number of keystrokes
- [guess the word currently being typed on the basis of:
 - the part of the word typed so far (can be no letters)
 - a language model (tells the likelihood of every word given the previous few words and possibly other inputs)

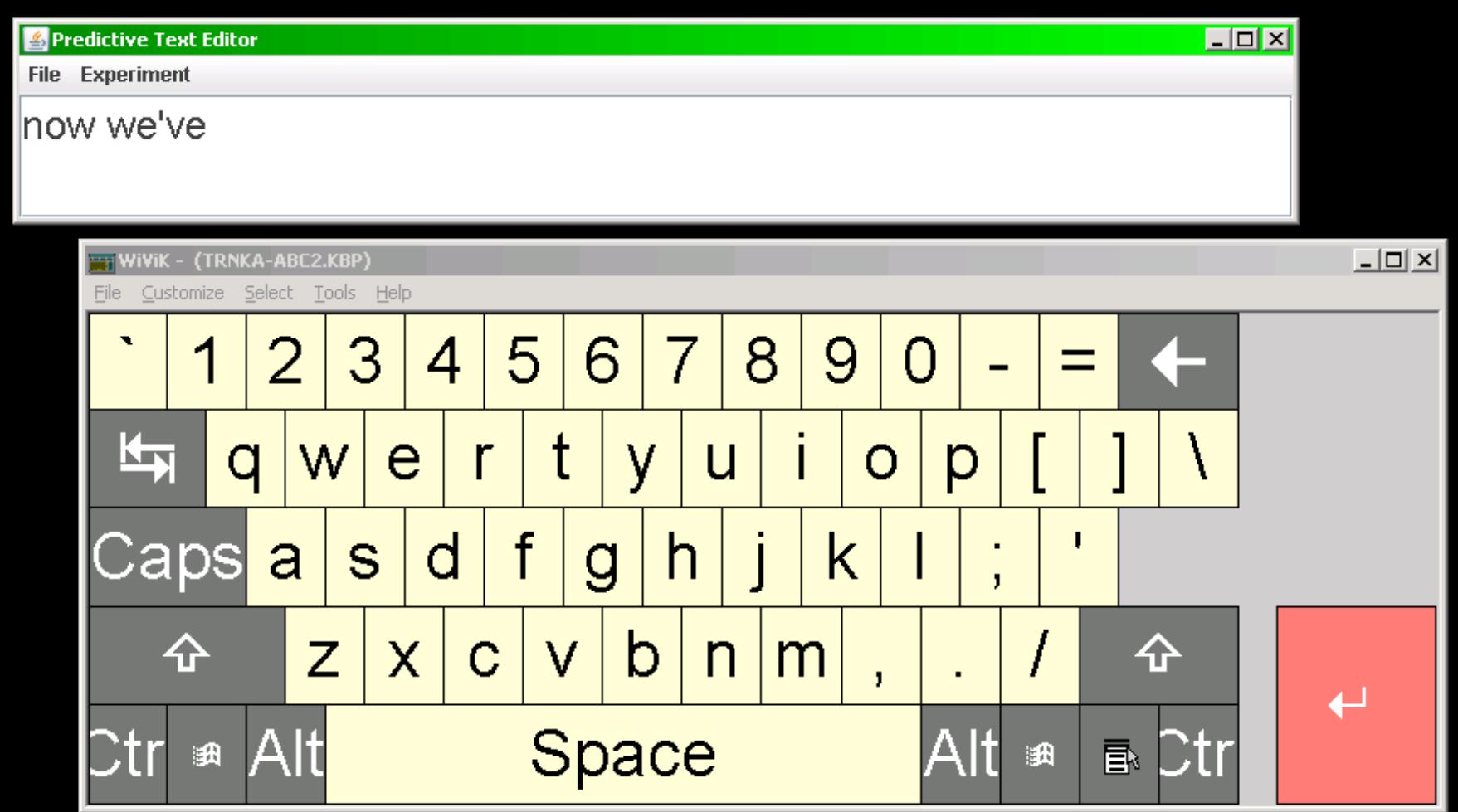
AAC Background



AAC Background



AAC Background



Predicting Words

Steps

- Filter the vocabulary by the prefix
- Compute the probability of all matching words given the context (previous words)
- Sort the list
- Present the top W words in order

Language Modeling

- [Language models provide the probability of a word in context
- [Trigram models are a typical language model, focusing on previous two words:

$$P(w \mid w_{-1}, w_{-2})$$

Language Modeling

- [Problem: where do the probabilities come from?
 - train the model by estimating the probabilities from a collection of text (corpus)

- [Problem: how do we evaluate the system?
 - measure keystroke savings on text not used in training (called testing)

Evaluating Word Prediction

- Keystroke savings on testing text

$$KS = \frac{chars - keystrokes}{chars} \times 100\%$$

- Simulated ideal user

- Simulated user interface – 5 predictions

Problem Statement

— [Training corpus generates probabilities that are going to work best on similar text

— but no appropriate AAC corpora exist!

— [Subproblems

— What should we use for training?

— How should we do testing?

Solving the Corpus Problem

- [Select a variety of corpora to reflect actual usage
 - focus on spoken language, but also written language, emails, etc.
- [Transform text to be more AAC-like
 - remove speech repairs from spoken texts
- [Build a small AAC corpus

Spoken Corpora

- [**Switchboard (2.8M words)** – telephone conversations that are centered on specific topics]
- [**Micase (545K words)** – university-setting conversation]
- [**SBCSAE (237K words)** – mostly face-to-face conversation]
- [**Charlotte (188K words)** – speech around Charlotte, NC]
- [**Callhome (48K words)** – telephone conversations between friends and family]

Other Corpora

- [AAC Email Corpus (28K words) – public mailing list archive, filtered by AAC users
- [Slate Magazine (4.2M words) – online magazine, similar to newspaper style

How to use our corpora?

[How should we test on the corpora?

[How should we train the model?

— In-domain – training on the same corpus

— Out-of-domain – trained on the other corpora

— Mixed-domain – training mixes in-domain and out-of-domain training

Domain Variations

- [In-domain – training and testing on the same corpus
 - most evaluations
- [Out-of-domain – testing corpus is not used in training
 - sometimes used to validate improvements
- [Mixed-domain – training mixes in-domain and out-of-domain
 - similar to an adaptive language model

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	red	green	green	green	green
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	green	red	green	green	green
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	green	green	red	green	green
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	green	green	green	red	green
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	green	green	green	green	red
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

In-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A					
Corpus B	red	green	green	green	green
Corpus C					
Corpus D					

green = training, red = testing

Domain Variations

- [In-domain – training and testing on the same corpus
 - most evaluations
- [Out-of-domain – testing corpus is not used in training
 - sometimes used to validate improvements
- [Mixed-domain – training mixes in-domain and out-of-domain
 - similar to an adaptive language model

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	red				
Corpus B		green	green	green	green
Corpus C		green	green	green	green
Corpus D		green	green	green	green

green = training, red = testing

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A		red			
Corpus B	green		green	green	green
Corpus C	green		green	green	green
Corpus D	green		green	green	green

green = training, red = testing

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A					
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A					
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A					Red
Corpus B	Green	Green	Green	Green	
Corpus C	Green	Green	Green	Green	
Corpus D	Green	Green	Green	Green	

green = training, red = testing

Out-of-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A					
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

Domain Variations

- [In-domain – training and testing on the same corpus
 - most evaluations
- [Out-of-domain – testing corpus is not used in training
 - sometimes used to validate improvements
- [Mixed-domain – training mixes in-domain and out-of-domain
 - similar to an adaptive language model

Mixed-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	red	green	green	green	green
Corpus B	white	green	green	green	green
Corpus C	white	green	green	green	green
Corpus D	white	green	green	green	green

green = training, red = testing

Mixed-domain Example

	Set 1	Set 2	Set 3	Set 4	Set 5
Corpus A	green	red	green	green	green
Corpus B	green	white	green	green	green
Corpus C	green	white	green	green	green
Corpus D	green	white	green	green	green

green = training, red = testing

Trigram Model Evaluation

Keystroke savings

↓
size

Testing corpus	In-domain training
AAC Email	48.92%
Callhome	43.76%
Charlotte	48.30%
SBCSAE	42.30%
Micase	49.00%
Switchboard	60.35%
Slate	53.13%

Trigram Model Evaluation

Keystroke savings



Testing corpus	In-domain training
AAC Email	48.92%
Callhome	43.76%
Charlotte	48.30%
SBCSAE	42.30%
Micase	49.00%
Switchboard	60.35%
Slate	53.13%

Trigram Model Evaluation

Keystroke savings



Testing corpus	In-domain training
AAC Email	48.92%
Callhome	43.76%
Charlotte	48.30%
SBCSAE	42.30%
Micase	49.00%
Switchboard	60.35%
Slate	53.13%

Trigram Model Evaluation

Keystroke savings

↓
size

Testing corpus	In-domain training
AAC Email	48.92%
Callhome	43.76%
Charlotte	48.30%
SBCSAE	42.30%
Micase	49.00%
Switchboard	60.35%
Slate	53.13%

Trigram Model Evaluation

Keystroke savings

	Training text	
Testing corpus	In-domain	Out-of-domain
AAC Email	48.92%	47.89%
Callhome	43.76%	52.95%
Charlotte	48.30%	52.44%
SBCSAE	42.30%	46.97%
Micase	49.00%	49.62%
Switchboard	60.35%	53.88%
Slate	53.13%	40.73%

Trigram Model Evaluation

Keystroke savings

	Training text	
Testing corpus	In-domain	Out-of-domain
AAC Email	48.92%	47.89%
Callhome	43.76%	52.95%
Charlotte	48.30%	52.44%
SBCSAE	42.30%	46.97%
Micase	49.00%	49.62%
Switchboard	60.35%	53.88%
Slate	53.13%	40.73%

Trigram Model Evaluation

Keystroke savings

	Training text	
Testing corpus	In-domain	Out-of-domain
AAC Email	48.92%	47.89%
Callhome	43.76%	52.95%
Charlotte	48.30%	52.44%
SBCSAE	42.30%	46.97%
Micase	49.00%	49.62%
Switchboard	60.35%	53.88%
Slate	53.13%	40.73%

Trigram Model Evaluation

Keystroke savings

	Training text	
Testing corpus	In-domain	Out-of-domain
AAC Email	48.92%	47.89%
Callhome	43.76%	52.95%
Charlotte	48.30%	52.44%
SBCSAE	42.30%	46.97%
Micase	49.00%	49.62%
Switchboard	60.35%	53.88%
Slate	53.13%	40.73%

Trigram Model Evaluation

Keystroke savings

	Training text		
Testing corpus	In-domain	Out-of-domain	Mixed-domain
AAC Email	48.92%	47.89%	52.18%
Callhome	43.76%	52.95%	53.14%
Charlotte	48.30%	52.44%	53.50%
SBCSAE	42.30%	46.97%	47.78%
Micase	49.00%	49.62%	51.46%
Switchboard	60.35%	53.88%	59.80%
Slate	53.13%	40.73%	53.05%

Trigram Model Evaluation

Keystroke savings

Testing corpus	Training text		
	In-domain	Out-of-domain	Mixed-domain
AAC Email	48.92%	47.89%	52.18%
Callhome	43.76%	52.95%	53.14%
Charlotte	48.30%	52.44%	53.50%
SBCSAE	42.30%	46.97%	47.78%
Micase	49.00%	49.62%	51.46%
Switchboard	60.35%	53.88%	59.80%
Slate	53.13%	40.73%	53.05%

Trigram Model Evaluation

Keystroke savings

	Training text		
Testing corpus	In-domain	Out-of-domain	Mixed-domain
AAC Email	48.92%	47.89%	52.18%
Callhome	43.76%	52.95%	53.14%
Charlotte	48.30%	52.44%	53.50%
SBCSAE	42.30%	46.97%	47.78%
Micase	49.00%	49.62%	51.46%
Switchboard	60.35%	53.88%	59.80%
Slate	53.13%	40.73%	53.05%

Topic Modeling

- [Example of testing methods applied to an improvement
- [Goal: seamlessly adapt the predictions to the topic
 - Build a separate trigram model for each topic in Switchboard
 - Combine the topic models using a weighted average
 - Weights based on similarity to the conversation

Topic Model Evaluation

Keystroke savings

Testing corpus	Switchboard trigram	Switchboard topic
AAC Email	43.25%	43.53%
Callhome	49.33%	49.52%
Charlotte	49.64%	50.07%
SBCSAE	43.49%	43.90%
Micase	46.52%	46.99%
Switchboard	60.35%	61.48%
Slate	39.17%	39.78%

Summary

- Constructed a corpus suite to approximate AAC text
- In-domain vs. out-of-domain
- Mixed domain
- Out-of-domain topic modeling
- Future directions