Corpus Studies in Word Prediction

Keith Trnka and Kathleen F. McCoy
University of Delaware

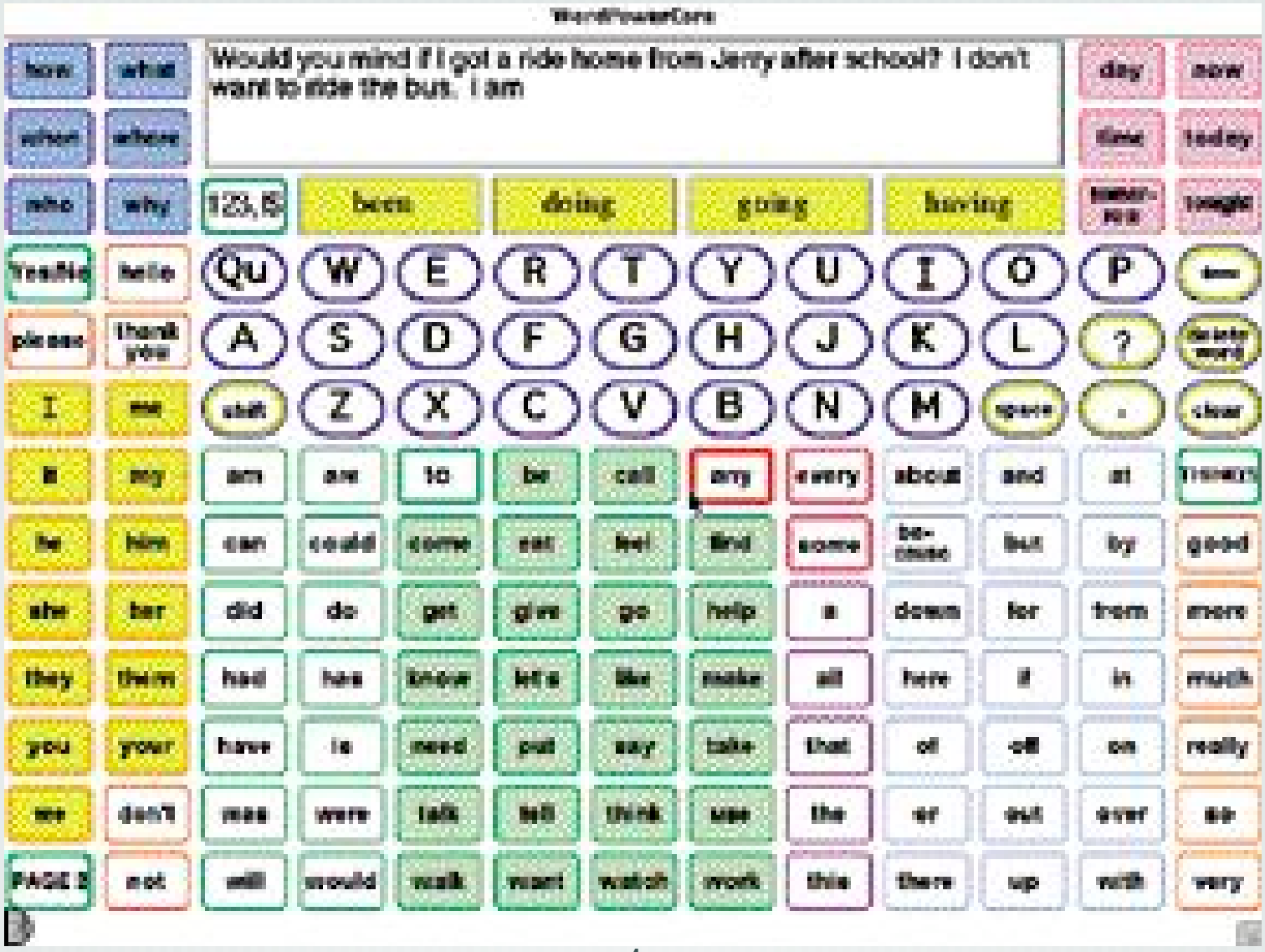
Outline

- 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/phase/icon input, speech synthesis output
- the communication rate divide and fatigue

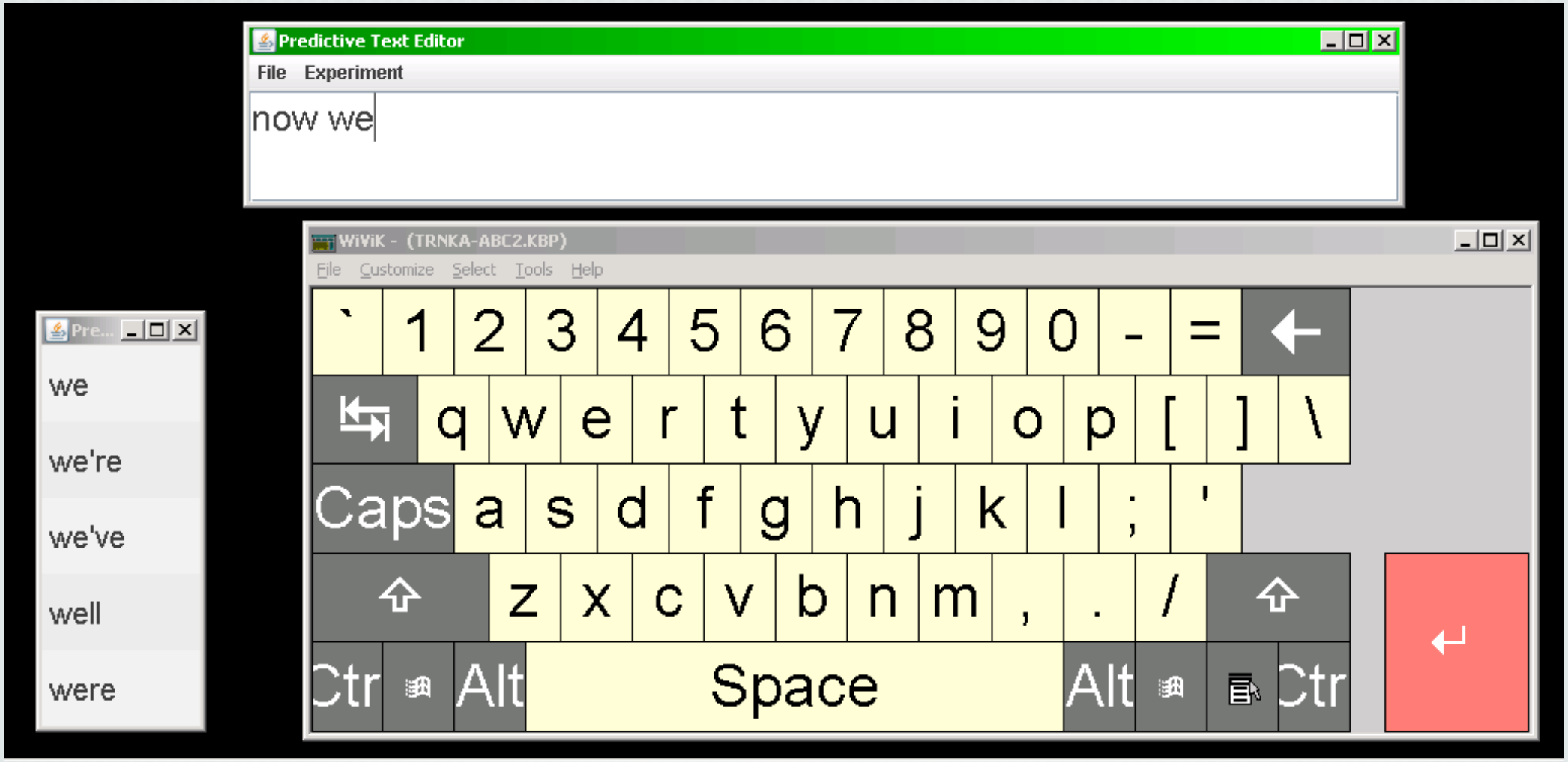
Example AAC Device



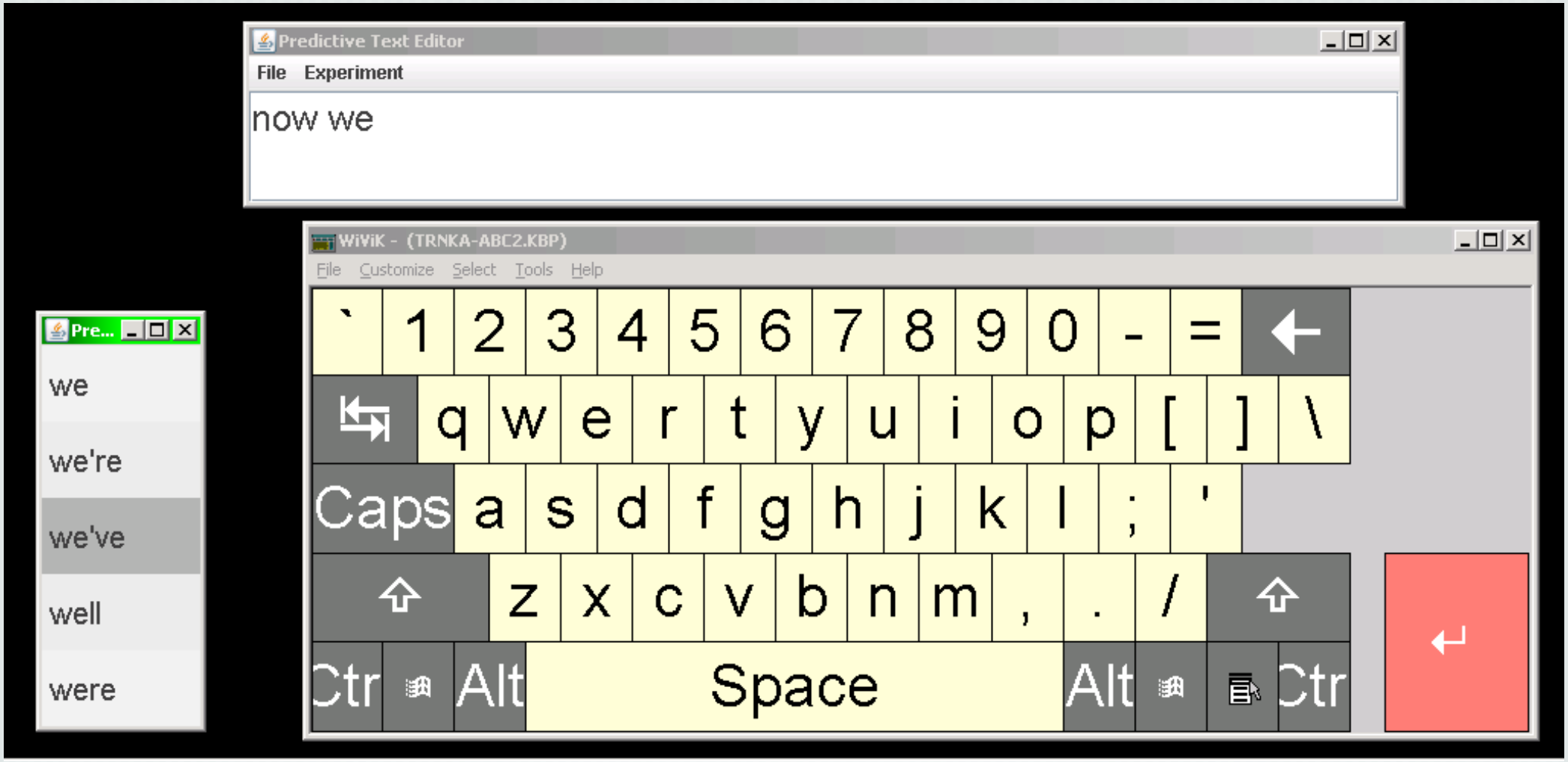
Word Prediction in AAC

- 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)

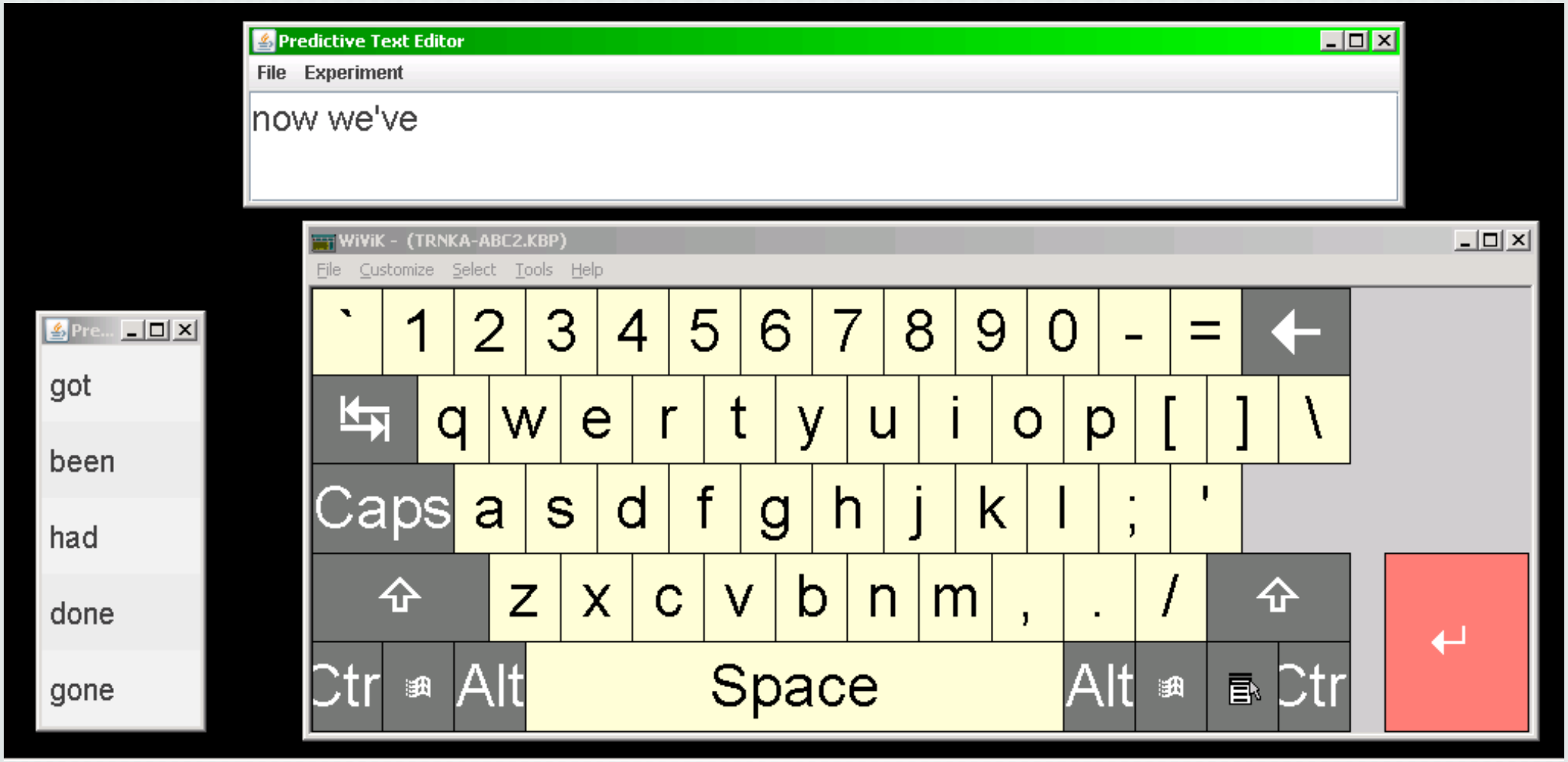
AAC Background



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					
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					
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					
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					
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					
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					
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					
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					
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					
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

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					
Corpus B					
Corpus C					
Corpus D					

green = training, red = testing

Mixed-domain Example

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

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

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

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

	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

	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

	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%**

statistically significant

Summary

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