GenERRate: Generating Errors for Use in Grammatical Error Detection

Jennifer Foster and Øistein E. Andersen

NAACL Workshop on Innovative Uses of NLP
for Building Educational Applications
Boulder, Colorado — 5th June 2009

Thanks to: Cambridge ESOL Examinations; CUP;
James Hunter, Gonzaga College

Talk Overview

Generating Ungrammatical Language for Use in Grammatical Error Detection

- **Why** it might be useful
- **When** it has been used before
- **How** it can be done
  - GenERRate
  - Testing GenERRate
    - Spoken Language Learner Corpus
    - Cambridge Learner Corpus

Talk Overview

Generating Ungrammatical Language for Use in Grammatical Error Detection

- **Why** it might be useful
- **When** it has been used before
- **How** it can be done
  - GenERRate
  - Testing GenERRate
    - Spoken Language Learner Corpus
    - Cambridge Learner Corpus

Talk Overview

Generating Ungrammatical Language for Use in Grammatical Error Detection

- **Why** it might be useful
- **When** it has been used before
- **How** it can be done
  - GenERRate
  - Testing GenERRate
    - Spoken Language Learner Corpus
    - Cambridge Learner Corpus

Talk Overview

Generating Ungrammatical Language for Use in Grammatical Error Detection

- **Why** it might be useful
- **When** it has been used before
- **How** it can be done
  - GenERRate
  - Testing GenERRate
    - Spoken Language Learner Corpus
    - Cambridge Learner Corpus

Why

Two types of error detection

1. Targeted Error Detection
2. Grammaticality Rating

Why

Two types of evidence used in error detection

1. Positive: compare to some model of *normal* language
2. Negative: compare to some model of *deviant* language

Combining both types of evidence is likely to be useful.

Why

- Positive data is easy enough to obtain
- Negative data is less straightforward
  - Find the correct type of text
  - Annotate errors

A possible solution?

Create negative data *automatically*.

Why

- Cheap to create
  - Knowledge of errors still necessary
- The error will appear in *varied contexts* - useful for training
- Number and type of errors can be controlled

Why

- Cheap to create
  - Knowledge of errors still necessary
- The error will appear in *varied contexts* - useful for training
- Number and type of errors can be controlled

Why

- Cheap to create
  - Knowledge of errors still necessary
- The error will appear in *varied contexts* - useful for training
- Number and type of errors can be controlled

Why

- Cheap to create
  - Knowledge of errors still necessary
- The error will appear in *varied contexts* - useful for training
- Number and type of errors can be controlled

When

Targeted Error Detection

- Sjöbergh & Knutsson, 2005
- Brockett et al., 2006
- Lee & Seneff, 2008

When

Grammaticality Rating

- Wagner et al., 2007
- Okanahara & Tsujii, 2007

When

Robustness Evaluation

- Bigert et al., 2005
- Foster, 2007

When

Unsupervised Learning

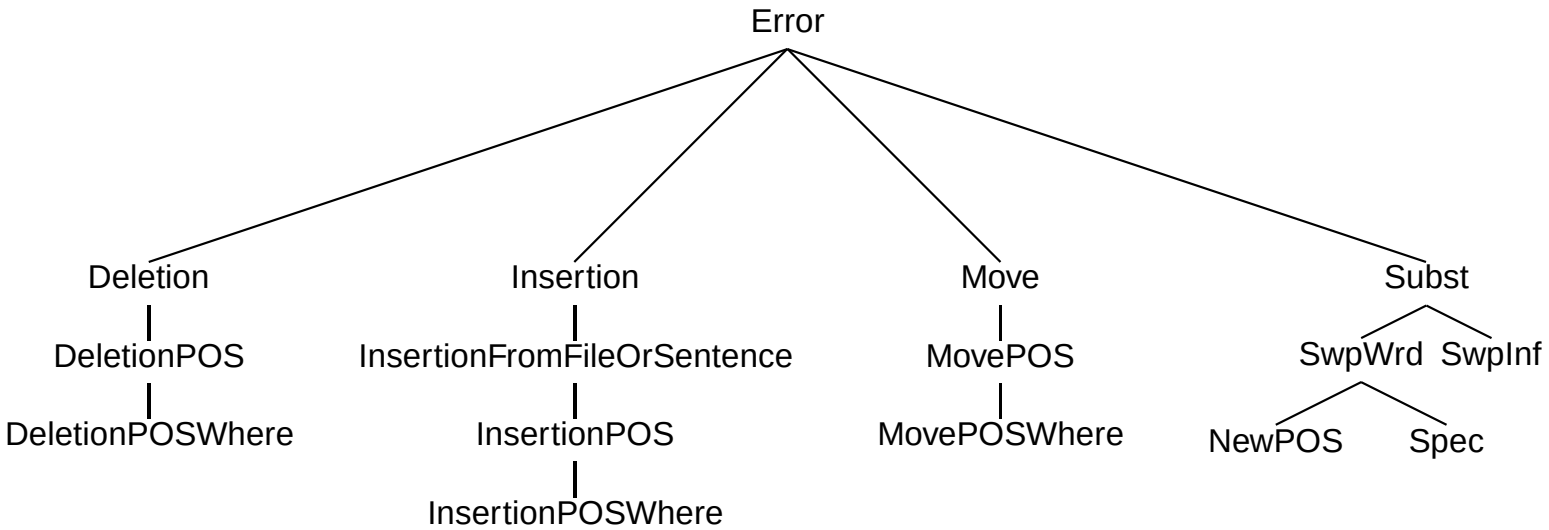
- Contrastive Estimation (Smith & Eisner, 2005)

How

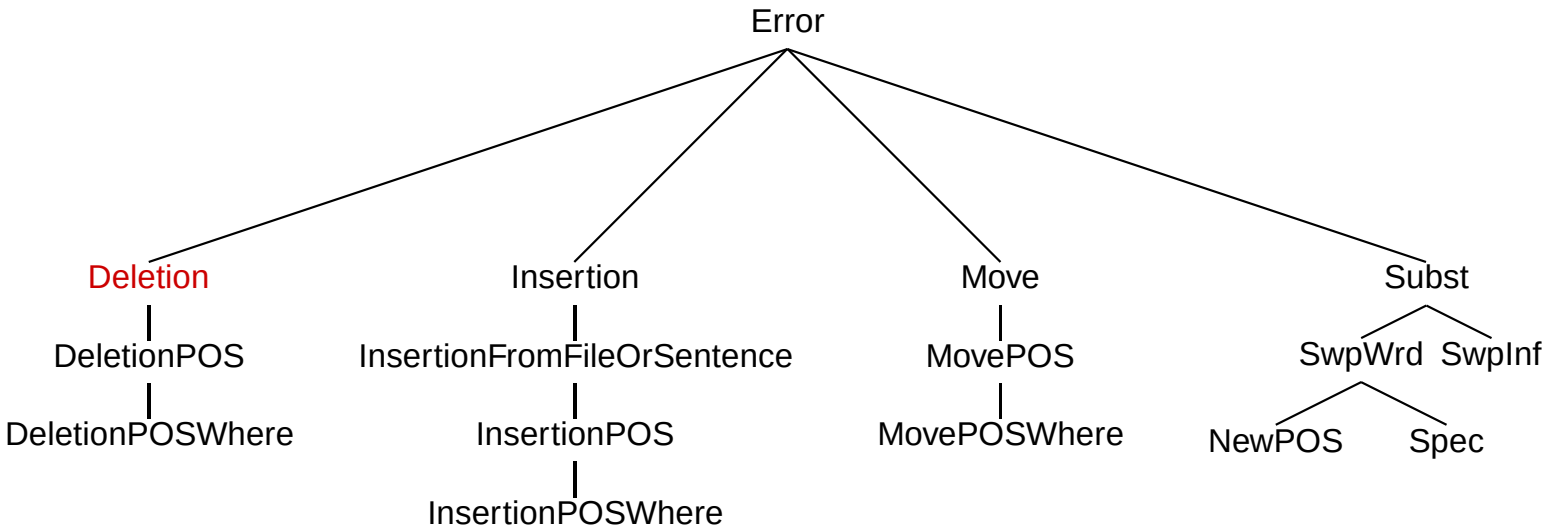
- GenERRate
- Available from

  `www.computing.dcu.ie/~jfoster/resources/generrate.html`

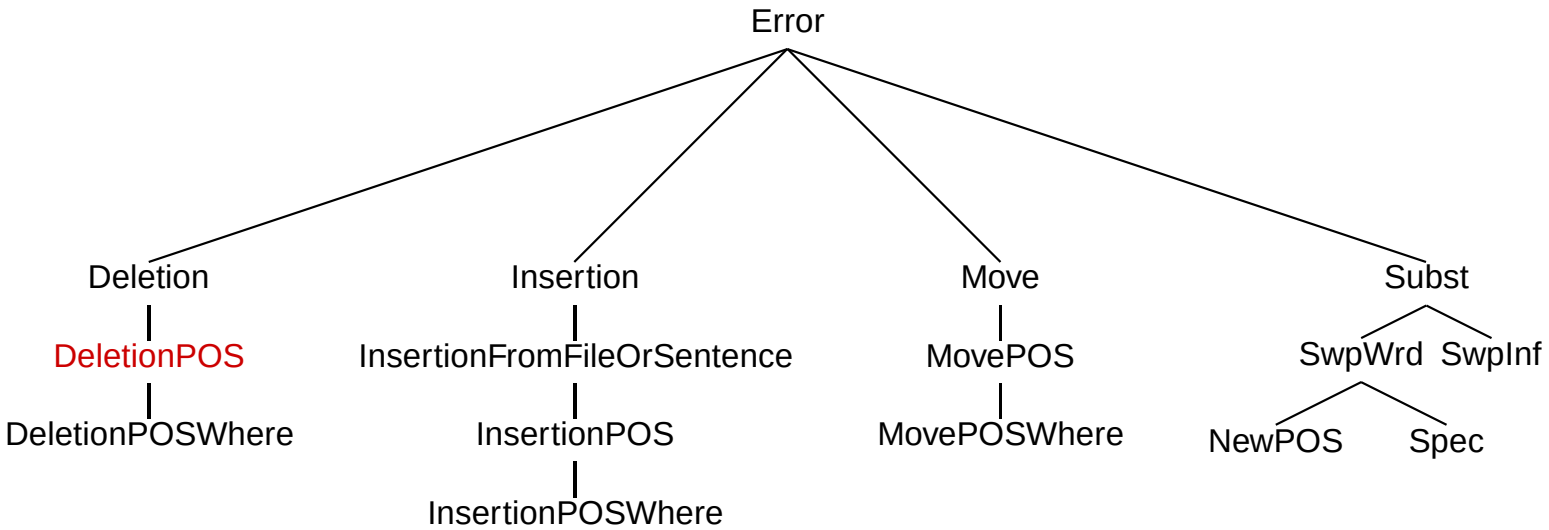
GenERRate: Supported Error Types



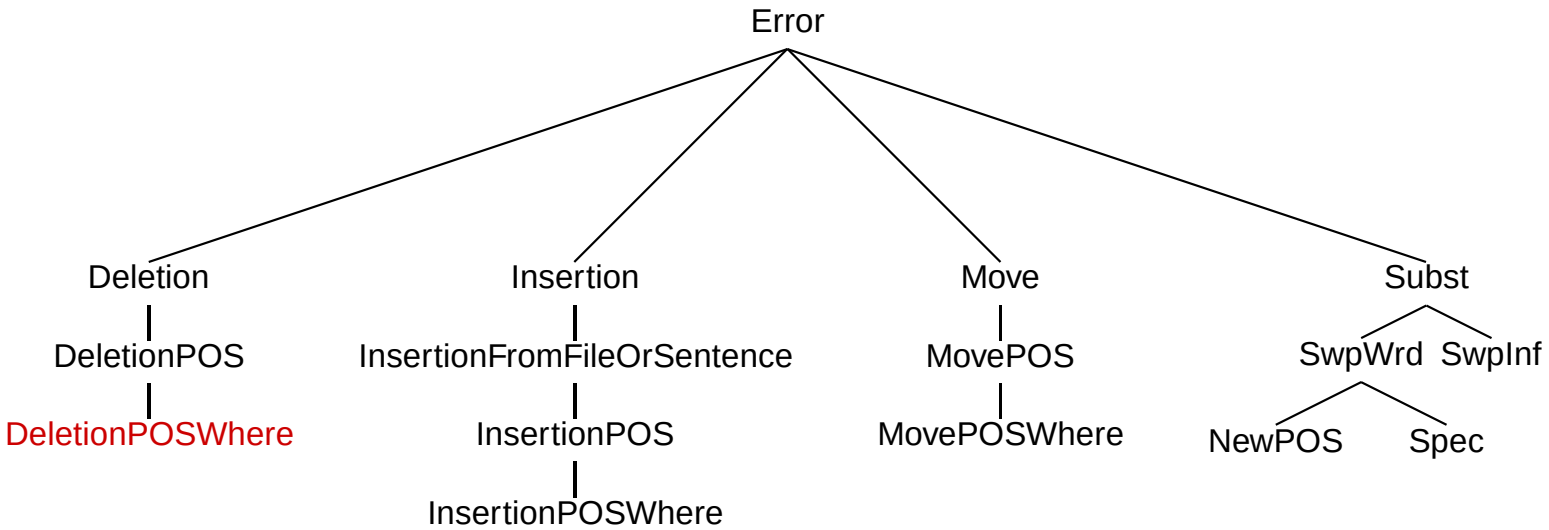
GenERRate: Supported Error Types



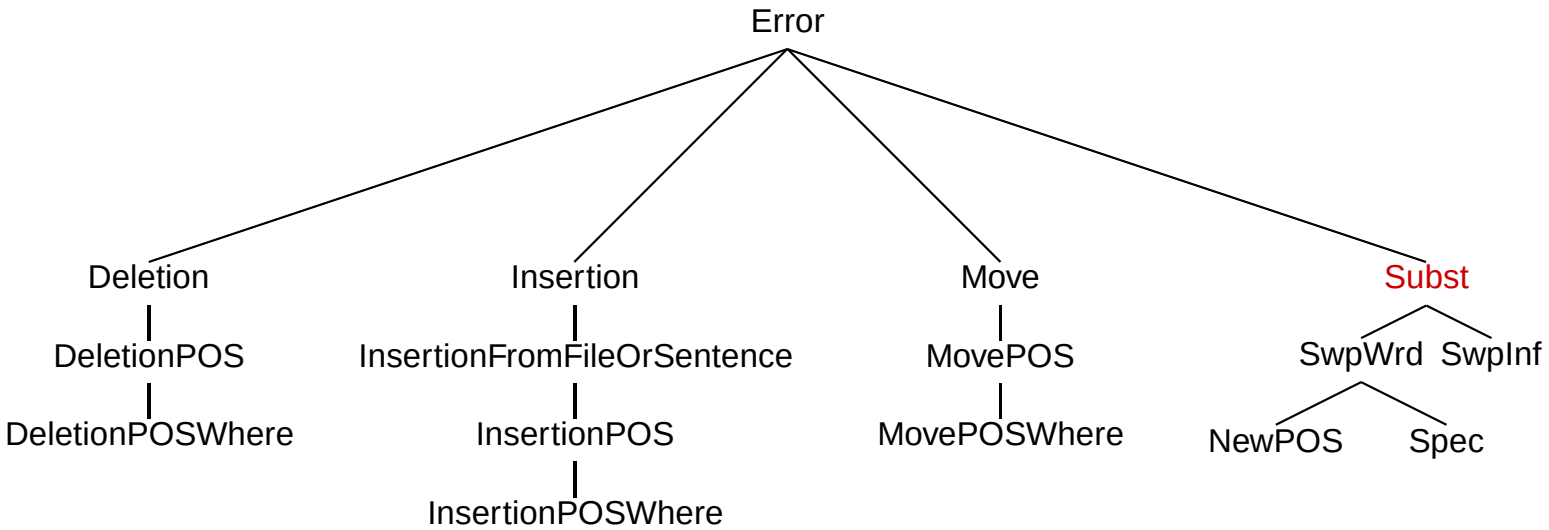
GenERRate: Supported Error Types



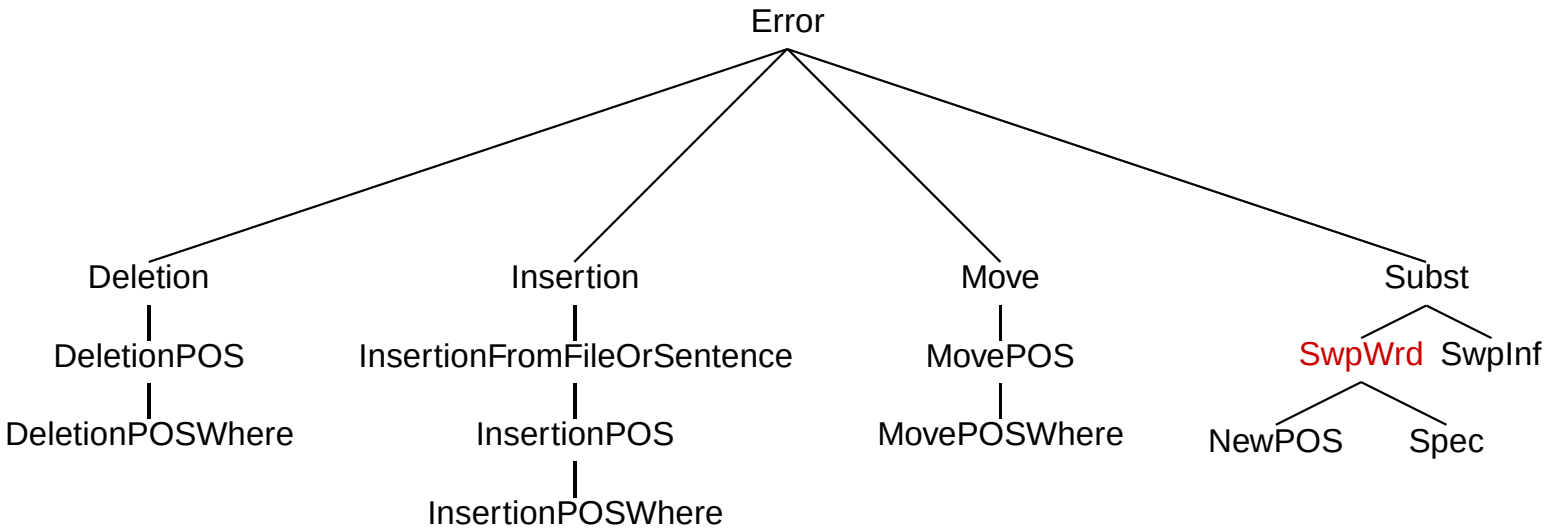
GenERRate: Supported Error Types



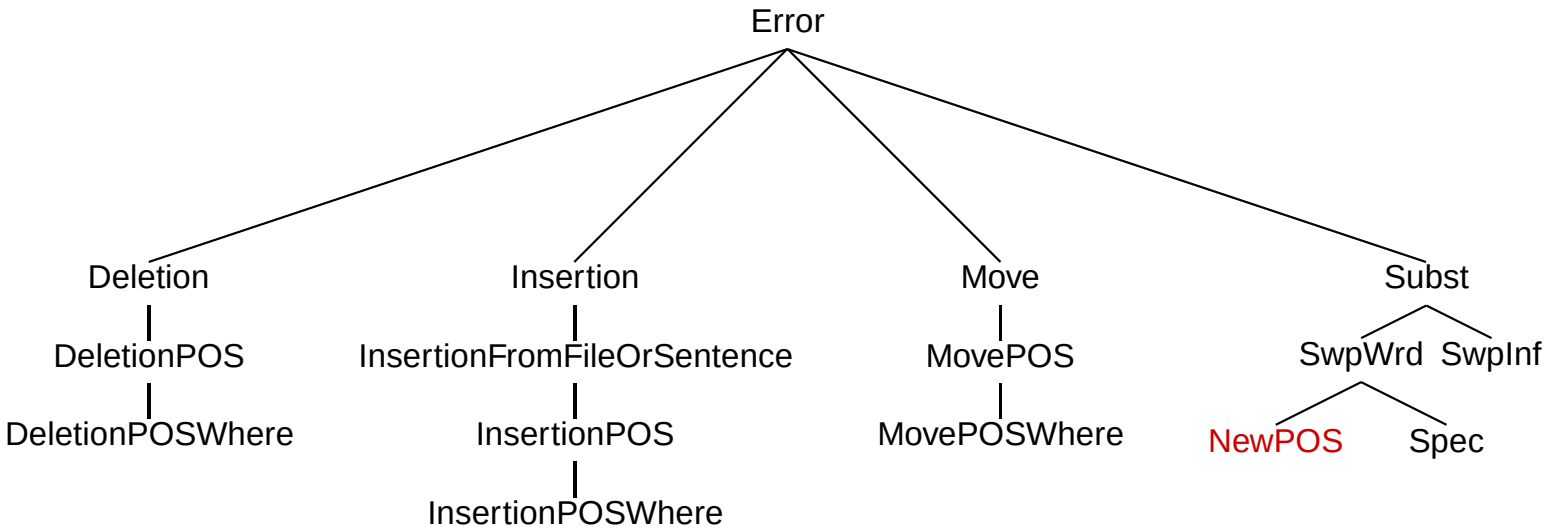
GenERRate: Supported Error Types



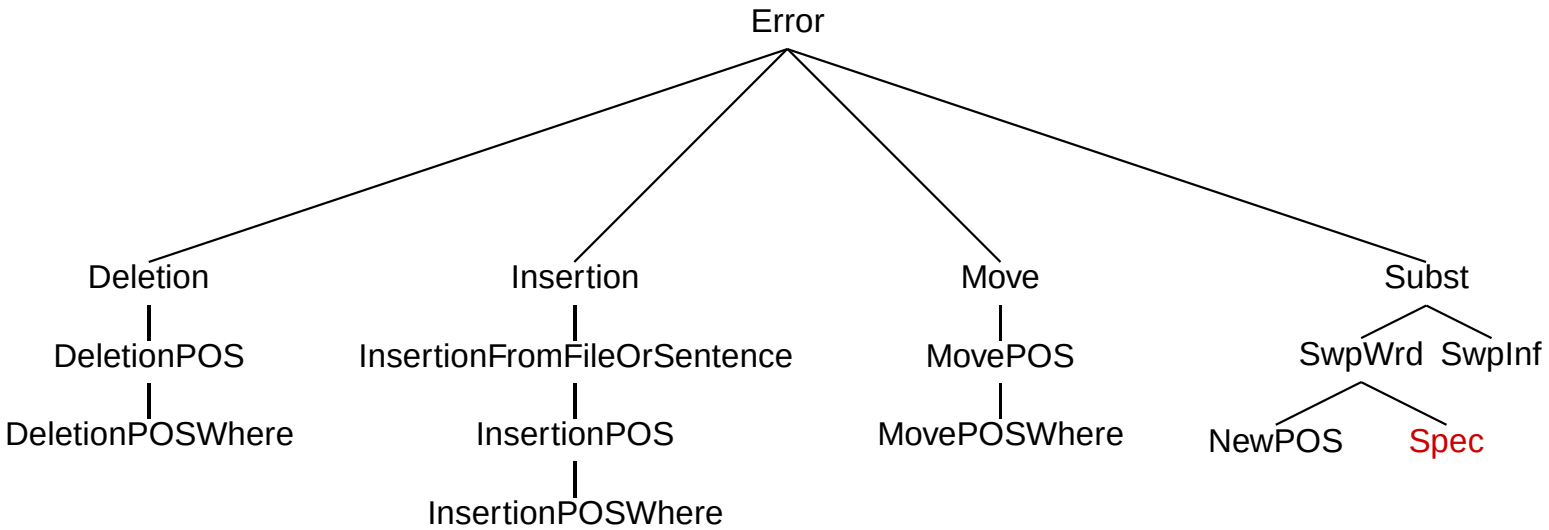
GenERRate: Supported Error Types



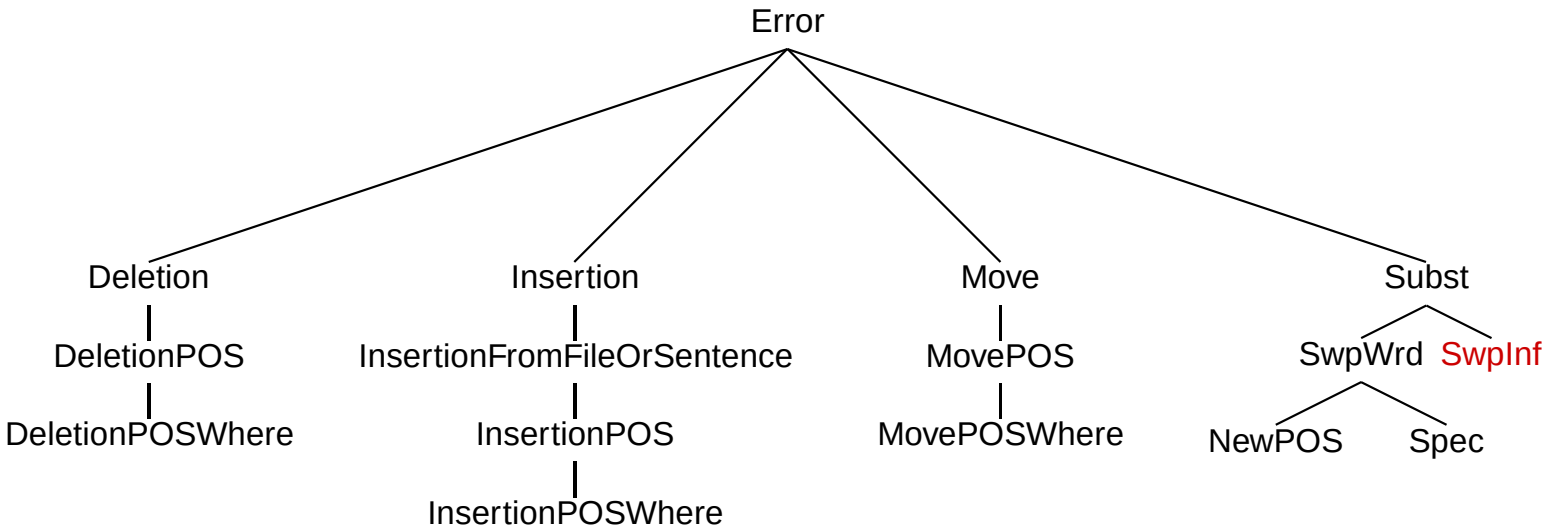
GenERRate: Supported Error Types



GenERRate: Supported Error Types



GenERRate: Supported Error Types



GenERRate

Input

1. A corpus of well-formed language
2. An error analysis file

Output

- An error-tagged corpus

GenERRate

Input

1. A corpus of well-formed language
2. An error analysis file

Output

- An error-tagged corpus

GenERRate

Error Analysis File

```
subst,word,an,a
subst,NNS,NN
subst,VBG,TO
delete,DT
move,RB,left,1
```

Input Corpus

The DT cats NNS are VBG also RB sitting VBG on IN the DT mat NN . .

Output Corpus

*The **cat are** also sitting on the mat .*
*The cats are also **to sit** on the mat .*
*The cats are also sitting **on mat** .*
*The cats **also are** sitting on the mat .*

GenERRate

Error Analysis File

```
subst,word,an,a
subst,NNS,NN
subst,VBG,TO
delete,DT
move,RB,left,1
```

Input Corpus

The DT cats NNS are VBG also RB sitting VBG on IN the DT mat NN . .

Output Corpus

*The **cat are** also sitting on the mat .*
*The cats are also **to sit** on the mat .*
*The cats are also sitting **on mat** .*
*The cats **also are** sitting on the mat .*

GenERRate

Error Analysis File

```
subst,word,an,a
subst,NNS,NN
subst,VBG,TO
delete,DT
move,RB,left,1
```

Input Corpus

The DT cats NNS are VBG also RB sitting VBG on IN the DT mat NN . .

Output Corpus

*The **cat are** also sitting on the mat .*
*The cats are also **to sit** on the mat .*
*The cats are also sitting **on mat** .*
*The cats **also are** sitting on the mat .*

Testing GenERRate

Two grammaticality rating experiments

- Distinguish ungrammatical sentences from learner corpora from corrected versions of these sentences
- Binary classification task

Spoken Learner Corpus Experiment

- Existing classifier that uses artificial data
- Can we improve the classifier by using *more realistic* training data?

Spoken Learner Corpus Experiment

The existing classifier

- Wagner et al., 2007
- *n*-gram frequency counts
- Training data
  - BNC sentences
  - Distorted versions of the BNC sentences
- Test data
  - Sentences from a spoken language learner corpus

Spoken Learner Corpus Experiment

The Spoken Learner Corpus

- 4,295 utterances
- Produced by ESL learners in a classroom setting
- Various levels and L1s
- Transcribed by the teacher
- Approx. 500 of these have been corrected

Spoken Learner Corpus Experiment

The new classifier

1. Take out 200 sentences from test data
2. Perform manual error analysis
3. Produce GenERRate error analysis file
4. Use GenERRate to generate new ungrammatical training data

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Biogas production is growing rapidly*

- **OLD**: *Biogas **production production** is growing rapidly*
- **NEW**: *Biogas **productions** is growing rapidly*

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Biogas production is growing rapidly*

- **OLD**: *Biogas **production production** is growing rapidly*
- **NEW**: *Biogas **productions** is growing rapidly*

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Biogas production is growing rapidly*

- **OLD**: *Biogas **production production** is growing rapidly*
- **NEW**: *Biogas **productions** is growing rapidly*

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Emil was courteous and helpful*

- **OLD**: *Emil **as** courteous and helpful*
- **NEW**: *Emil courteous and **was** helpful*

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Emil was courteous and helpful*

- **OLD**: *Emil **as** courteous and helpful*
- **NEW**: *Emil courteous and **was** helpful*

Spoken Learner Corpus Experiment

Training data examples

ORIGINAL: *Emil was courteous and helpful*

- **OLD**: *Emil **as** courteous and helpful*
- **NEW**: *Emil courteous and **was** helpful*

Spoken Learner Corpus Experiment

Results

- OLD CLASSIFIER
  37.0% of the ungrammatical sentences are flagged and **95.5%** of the flagged sentences are ungrammatical.
- NEW CLASSIFIER
  51.6% of the ungrammatical sentences are flagged and **94.9%** of the flagged sentences are ungrammatical.

Spoken Learner Corpus Experiment

Results

- OLD CLASSIFIER
  37.0% of the ungrammatical sentences are flagged and **95.5%** of the flagged sentences are ungrammatical.
- NEW CLASSIFIER
  51.6% of the ungrammatical sentences are flagged and **94.9%** of the flagged sentences are ungrammatical.

Spoken Learner Corpus Experiment

Results

- OLD CLASSIFIER
  37.0% of the ungrammatical sentences are flagged and **95.5%** of the flagged sentences are ungrammatical.
- NEW CLASSIFIER
  51.6% of the ungrammatical sentences are flagged and **94.9%** of the flagged sentences are ungrammatical.

Cambridge Learner Corpus Experiment

- Create a large error corpus inspired by the Cambridge Learner Corpus
- Compare a classifier trained on the artificial corpus to one trained on the CLC

Cambridge Learner Corpus Experiment

The CLC

- Approx. 30 million words of Learner English
- Collected from University of Cambridge ESOL papers
- Variety of learner levels
- Approx. 50% annotated for errors and corrected

Cambridge Learner Corpus Experiment

The existing classifier

- Andersen 2006
- POS and token *n*-grams
- Training and test data
  - CLC sentences
  - corrections of CLC sentences

Cambridge Learner Corpus Experiment

The new classifier

Trained on artificial data, generated from corrected CLC sentences using GenERRate

- Extract POS trigrams from error-annotated CLC sentences
- Convert POS trigrams to errors supported by GenERRate
- CLC error analysis file

Cambridge Learner Corpus Experiment

The new classifier

Trained on artificial data, generated from corrected CLC sentences using GenERRate

- Extract POS trigrams from error-annotated CLC sentences
- Convert POS trigrams to errors supported by GenERRate
- CLC error analysis file

Cambridge Learner Corpus Experiment

Results

- OLD CLASSIFIER
  Flags **42.6%** of the ungrammatical sentences and **69.7%** of flagged cases are ungrammatical
- NEW CLASSIFIER
  Flags **30.7%** of the ungrammatical sentences and **62%** of flagged cases are ungrammatical

Cambridge Learner Corpus Experiment

Results

- OLD CLASSIFIER
  Flags **42.6%** of the ungrammatical sentences and **69.7%** of flagged cases are ungrammatical
- NEW CLASSIFIER
  Flags **30.7%** of the ungrammatical sentences and **62%** of flagged cases are ungrammatical

Cambridge Learner Corpus Experiment

Results

- OLD CLASSIFIER
  Flags **42.6%** of the ungrammatical sentences and **69.7%** of flagged cases are ungrammatical
- NEW CLASSIFIER
  Flags **30.7%** of the ungrammatical sentences and **62%** of flagged cases are ungrammatical

Cambridge Learner Corpus Experiment

Reasons for accuracy drop

- Some errors simply not supported by GenERRate
  - spelling mistakes
  - more complicated errors (*a other woman* $\rightarrow$ *another woman*)
- Multiple errors per sentence in real data
- POS tagset not fine-grained enough, e.g. mass vs count nouns

Cambridge Learner Corpus Experiment

Reasons for accuracy drop

- Some errors simply not supported by GenERRate
  - spelling mistakes
  - more complicated errors (*a other woman* $\rightarrow$ *another woman*)
- Multiple errors per sentence in real data
- POS tagset not fine-grained enough, e.g. mass vs count nouns

Cambridge Learner Corpus Experiment

Reasons for accuracy drop

- Some errors simply not supported by GenERRate
  - spelling mistakes
  - more complicated errors (*a other woman* $\rightarrow$ *another woman*)
- Multiple errors per sentence in real data
- POS tagset not fine-grained enough, e.g. mass vs count nouns

The Problem of Covert Errors

When to avoid them

- *The cats are worth seeing.*

  $\rightarrow$

  *The cats are worth **to see**.*

- *The cats are also sitting on the mat.*

  $\rightarrow$

  *The cats are also **to sit** on the mat.*

The Problem of Covert Errors

When to avoid them

- *The cats are worth seeing.*

  $\longrightarrow$

  *The cats are worth **to see**.*

- *The cats are also sitting on the mat.*

  $\longrightarrow$

  *The cats are also **to sit** on the mat.*

The Problem of Covert Errors

When to avoid them

- *The cats are worth seeing.*

  $\longrightarrow$

  *The cats are worth **to see**.*

- *The cats are also sitting on the mat.*

  $\longrightarrow$

  *The cats are also **to sit** on the mat.*

The Problem of Covert Errors

When to avoid them

- *The cats are worth seeing.*

  $\longrightarrow$

  *The cats are worth **to see**.*
- *The cats are also sitting on the mat.*

  $\longrightarrow$

  *The cats are also **to sit** on the mat.*

The Problem of Covert Errors

When to avoid them

- *The cats are worth seeing.*

  $\longrightarrow$

  *The cats are worth **to see**.*
- *The cats are also sitting on the mat.*

  $\longrightarrow$

  *The cats are also **to sit** on the mat.*

The Problem of Covert Errors

When **not** to avoid them

- *—What time did you go to bed in high school?*
  — I went to bed at one.
  →
  —What time did you go to bed in high school?
  *— I **go** to bed at one.*
- *When I was a high school student I went to bed at one in the morning*
  →
  *When I was a high school student I **go** to bed at one in the morning*

The Problem of Covert Errors

When **not** to avoid them

- *—What time did you go to bed in high school?*
  — I went to bed at one.
  →
  —What time did you go to bed in high school?
  *— I **go** to bed at one.*
- *When I was a high school student I went to bed at one in the morning*
  →
  *When I was a high school student I **go** to bed at one in the morning*

The Problem of Covert Errors

When **not** to avoid them

- *—What time did you go to bed in high school?*
  — I went to bed at one.
  →
  —What time did you go to bed in high school?
  *— I **go** to bed at one.*
- *When I was a high school student I went to bed at one in the morning*
  →
  *When I was a high school student I **go** to bed at one in the morning*

The Problem of Covert Errors

When **not** to avoid them

- *—What time did you go to bed in high school?*
  — I went to bed at one.
  $\rightarrow$
  —What time did you go to bed in high school?
  *— I **go** to bed at one.*
- *When I was a high school student I went to bed at one in the morning*
  $\rightarrow$
  *When I was a high school student I **go** to bed at one in the morning*

The Problem of Covert Errors

When **not** to avoid them

- *—What time did you go to bed in high school?*
  — I went to bed at one.
  $\rightarrow$
  —What time did you go to bed in high school?
  *— I **go** to bed at one.*
- *When I was a high school student I went to bed at one in the morning*
  $\rightarrow$
  *When I was a high school student I **go** to bed at one in the morning*

Next version of GenERRate

- Integration with WordNet
- Spelling errors
- Different ways of specifying contextual information, e.g. parsed input

Thanks for listening

Next version of GenERRate

- Integration with WordNet
- Spelling errors
- Different ways of specifying contextual information, e.g. parsed input

Thanks for listening

Postdoc opportunity in Dublin

- 18-month postdoc on a CALL project
- Dublin Institute of Technology and the National Digital Research Centre
- Experience in the following: information retrieval, machine translation, word sense disambiguation, topic categorisation.
- For more details contact: John.Kelleher@comp.dit.ie