Error tagging

Yukio Tono (TUFS)
Error tagging (1)

- Special type of annotation
- Normal procedure
  - Defining error types
  - Designing the tagset specification
  - Trial: error-tag sample data ➔ feedback
- Creating an error tag manual
  ➔ CLC
  ➔ ICLE
  ➔ SST
Oh. I see. Never been there. Eh, is that nice place?

Kind of.

Kind of.

Ye, yes. It's suburb.

Suburb.

But, beautiful place.

Mhm. I see. What did you there?

I studied. Yes, I entered the university one year as exchange student.
Error tagging schemes

- Depend on the research questions
- SST → surface grammatical/lexical errors

Error tagging format:

```
And, please describe this picture.</A>
Describe? <F>Mhm</F>. <.><./> Maybe, <SC>this</SC>
toda</SC> <F>mm</F> it is a sunny day, and
in front of <SC?>hou</SC> big house, <F>er</F> two
housewifes</n_inf>
talking each other.
```

<at odr="1" crr="a"> (article, order of correction, correct form)
# Aspects of errors in the schemes

<table>
<thead>
<tr>
<th></th>
<th>CLC</th>
<th>FreeText</th>
<th>Louvain</th>
<th>NICT JLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Punctuation</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Spelling</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Grammar</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Lexis</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>(register)</td>
<td>(register)</td>
<td>(register)</td>
<td>(non-verbal cues)</td>
</tr>
<tr>
<td>Discourse</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>(pronoun reference)</td>
<td>(cohesion)</td>
<td>(unnatural discourse)</td>
<td>(self-corrections)</td>
</tr>
</tbody>
</table>

Table 2. Levels of linguistic categorization.
Error taxonomy (James 1991)

- Linguistic category classification
- Surface taxonomy classification
  - Omission (*He'll pass his exam and I'll too. )
  - Addition (based on Dulay, Burt & Krashen)
    - Regularization: *buyed for bought
    - Irregularization: *dove for the past form dived.
    - Double marking: He doesn't know*s me.
  - Misformation (*I seen him yesterday. )
  - Misordering (* He every time comes late home. )
  - Blending (*according to Erica's opinion ➔ blending of according to Erica and in Erica's opinion)
FALKO system

i) POS,
ii) one or more corrections, and
iii) linguistic information at the levels orthography, word formation, agreement, government, tense, mood, word order and expression.

In turn, each of these linguistic levels breaks down further into three sub-levels associated with Corder’s (1974) and Ellis’ (1994) steps in EA:

i) identification of errors,
ii) description of errors, according to a target modification classification, and
iii) explanation of errors, i.e. source of errors,

Anke Lüdeling (Humboldt-Universität zu Berlin)
FRIDA corpus tagging system

Table 1: Error Domains and Categories

<table>
<thead>
<tr>
<th>Error Domains</th>
<th></th>
<th>Error Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;F&gt; Form</td>
<td>&lt;AGL&gt;</td>
<td>Agglutination</td>
</tr>
<tr>
<td></td>
<td>&lt;Maj&gt;</td>
<td>Upper/lower case</td>
</tr>
<tr>
<td></td>
<td>&lt;Dia&gt;</td>
<td>Diacritics</td>
</tr>
<tr>
<td></td>
<td>&lt;Hom&gt;</td>
<td>Homonymy</td>
</tr>
<tr>
<td></td>
<td>&lt;Gra&gt;</td>
<td>Other spelling errors</td>
</tr>
<tr>
<td>&lt;M&gt; Morphology</td>
<td>&lt;MDP&gt;</td>
<td>Derivation-prefixation</td>
</tr>
<tr>
<td></td>
<td>&lt;Mds&gt;</td>
<td>Derivation-suffixication</td>
</tr>
<tr>
<td></td>
<td>&lt;Mfl&gt;</td>
<td>Inflection</td>
</tr>
<tr>
<td></td>
<td>&lt;Mfc&gt;</td>
<td>Inflection-confusion</td>
</tr>
<tr>
<td></td>
<td>&lt;Mco&gt;</td>
<td>Compounding</td>
</tr>
<tr>
<td>&lt;G&gt; Grammar</td>
<td>&lt;Cla&gt;</td>
<td>Class</td>
</tr>
<tr>
<td></td>
<td>&lt;Aux&gt;</td>
<td>Auxiliary</td>
</tr>
<tr>
<td></td>
<td>&lt;Gen&gt;</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>&lt;Mod&gt;</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td>&lt;Nbr&gt;</td>
<td>Number</td>
</tr>
<tr>
<td></td>
<td>&lt;Per&gt;</td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td>&lt;Tps&gt;</td>
<td>Tense</td>
</tr>
<tr>
<td></td>
<td>&lt;Voi&gt;</td>
<td>Voice</td>
</tr>
<tr>
<td></td>
<td>&lt;Euf&gt;</td>
<td>Euphony</td>
</tr>
</tbody>
</table>

| L | Lexis       | SIG | Meaning                  |
|   | Adjective complementation |    | Adverb complementation   |
|   | Verb complementation      |    | Noun complementation     |
|   | Prefab                  |    |                         |
| X | Syntax       | ORD | Word order               |
|   | Word missing  |    |                         |
|   | Word redundant |     |                         |
|   | Cohesion      |    |                         |
| R | Register     | RLE | Lexis                    |
|   | RSY | Syntax      |    |                         |
| Y | Style        | CLR | Unclear                  |
|   | LOU | Heavy       |    |                         |
| Q | Punctuation  | CON | Punctuation confusion   |
|   | TRO | Punctuation redundant |     |                         |
|   | OUB | Punctuation missing       |    |                         |
| Z | Typo         |     |                         |
Cambridge Learner Corpus

i) wrong form
ii) missing
iii) word or phrase needs replacing
iv) word or phrase is unnecessary
v) word is wrongly derived

and a linguistic classification dimension centred around POS categorization, consisting of:

i) pronoun
ii) conjunction
iii) determiner
iv) adjective
v) noun
vi) quantifier
vii) preposition
viii) verb
ix) adverb.

The resulting annotation is shown in (9):

(9) […] lawyers, doctors, etc, <#UA>they</#UA> hardly earn #50,000 a year. (CLC; Nicholls 2003: 576)
Searching for different error types

- **Omission**
  - Search for all the occurrences of error tags which is followed by the closing tags immediately after the opening tags.
- ¥ is equal to “backslash” on Japanese PCs
Searching for different error types

- **Addition**
- Search for all the instances where learners put extra unnecessary words or phrases
Export the results to Excel

**STEPS**

1) Get the concordance lines
2) Choose [Display]-[Distribution]
3) Save the results
4) Open it from Excel as a text file (column format)
Export the results to Excel
Error Analysis (EA)

- Based on nativist views of language learning
  - Interlanguage (Selinker 1972)
  - Idiosyncratic dialect (Corder 1971)

- Basic steps:
  - Collection of a sample of learner language
  - Identification of errors
  - Description of errors
  - Explanation of errors
  - Error evaluation
Error descriptions 1

- Linguistic taxonomy:
  - Basic sentence structure
  - Verb phrase (tense/ aspect/ subjunctive/ auxiliary/ non-finite verb)
  - Verb complementation
  - Noun phrase
  - Prepositional phrase
  - Adjunct
  - Coordinate & subordinate constructions
  - Sentence connection
Error description 2

- **Surface structure (modification) taxonomy:**
  - Omission
  - Addition
    - Regularization: e.g. *eated for ate
    - Double-marking: e.g. He didn’t *came
    - Simple addition: e.g. regularization/double-marking 以外
  - Misinformation
    - Regularization: e.g. *Do they be happy? → Are they happy?
    - Archi-forms: e.g. It’s not me. Me don’t care.（両方me）
    - Alternating forms: e.g. Don’t watch. & No watch.
  - Misordering: e.g. She fights all the time her brother.
CL methods and LC

- Overuse vs. underuse

- Use vs. misuse (errors)
  - Linguistic classification of errors
    - Lexical vs. grammatical (POS + tense/agreement/etc)
  - Surface strategy taxonomy
    - Omissions/additions/ misinformations/ misorderings
      (Dulay, Burt & Krashen 1982)
SLA and CLR

- Description ➔ Explanation
- SLA theories:
  - UG-Based SLA (Hawkins, White) ➔ more focus on lexicon
  - Processibility Hypothesis (Pienemann) ➔ Levelt & LFG
  - Competition Model (MacWhinney) ➔ very much frequency-based
- Related disciplines:
  - Cognitive linguistics; Usage-based approach
  - Systemic-functional grammar
  - Natural language processing
  - Data mining; Neural network
Error freq’s & distributions
Dimension 1（64.2%） Dimension 2（寄与率20.6%）
Abe and Tono (2005)

Dimension 1 (65.5% of Inertia)

Dimension 2 (20.0% of Inertia)

Verb Noun

WR mode SP

error rate pattern

WR SP Low Verb High Noun
Error types across POS (Abe & Tono 2005)
Automatic error identification
Automatic identification of learner errors

- JEFLL Corpus → The error-corrected version is now ready.

- We are working on the program that can compare the original and corrected versions of the sentence and automatically identify the patterns of deviation from the corrected sentence in terms of the following 3 types of errors (James 1998):
  - Addition/ omission/ misformation
Parallel concordancing

Our school festival is on October 23rd and 24th. It had a...
Errors involved in copula “be”

Tokyo University of Foreign Studies

Friday, October 23rd and Saturday, October 24th were my school festival days.

I was in the on-stage performance club. (?)

Our class had a mosaic exhibition.

We sang “Kimi wo nosete” in the song contest.

Our class took part in the song and arts contests. (?)

I enjoyed the festival!

1C sang the song “Kimono-se.”

My class made a 3-D artwork. (?)

We exhibited a big artwork in our classroom. (?)

We sang a song called “Kimi wo nosete.”

The themes were flight, revolution, and nature.

Our school festival had collage and song contests.

The performance days were October twenty third, and twenty fourth.

Our school festival days were October 23rd and 24th.

My class sang “Kimi wo nosete.”

But neither 1-A, 1-B, 1-C nor 1-D won any prizes. (?)

Many club’s gave a performance.
DP matching

INPUT (corrected):

\[ W-a \rightarrow W-b \rightarrow W-c \rightarrow \ldots \rightarrow W-i \]

INPUT (original):

\[ W-a \rightarrow W-b' \rightarrow W-d \rightarrow \ldots \rightarrow W-i \]

ANALYSIS:

\[ W-a \rightarrow W-b \rightarrow W-c \rightarrow W-d \rightarrow \ldots \rightarrow W-i \]

[msf] [oms] [add]
Automatic identification of learner errors

The first reason is every member of my family is busy in the morning.

- Looking at n-grams for maximum match and analyse the unmatched elements:
Automatic identification: output

T: My mother cooks very well ← corrected sentence
O: mother is cook very well ← original sentence
A: <oms>My</oms> mother <add>is</add> cook[*]:msf very well ← identifying differences

☐ Correspondence ratio:
  ■ Word level: 3/5
  ■ Character level: 3.80/5(76%)

Notes:  T = target; O = original; A = analysis
Looking for criterial features
Distributions of error types

- S3
- S2
- S1
- J3
- J2
- J1

Legend:
- Omission
- Addition
Distributions of error types

- S3
- S2
- S1
- J3
- J2
- J1

Legend:
- Omission
- Addition
Omission errors are significantly more frequent than addition errors.

Errors will not decrease across proficiency.
More nominalization at advanced level, which increases the number of “of-addition/omission” errors.
Use of modals

Marked omissions at the very beginning stages
Later, more use of modals lead to more addition errors
Errors related to ‘have’

N-word clusters of “have”

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>515</td>
<td>have &lt;cons</td>
</tr>
<tr>
<td>2</td>
<td>388</td>
<td>have &lt;add</td>
</tr>
<tr>
<td>3</td>
<td>303</td>
<td>I have</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>had &lt;cons</td>
</tr>
<tr>
<td>5</td>
<td>218</td>
<td>have &lt;add=a</td>
</tr>
<tr>
<td>6</td>
<td>210</td>
<td>have &lt;add=a&lt;/add&gt;</td>
</tr>
<tr>
<td>7</td>
<td>213</td>
<td>have &lt;cons=a&lt;/cons&gt;</td>
</tr>
<tr>
<td>8</td>
<td>213</td>
<td>have &lt;cons=a&lt;/cons&gt;</td>
</tr>
<tr>
<td>9</td>
<td>167</td>
<td>I have</td>
</tr>
<tr>
<td>10</td>
<td>153</td>
<td>had &lt;cons=a</td>
</tr>
<tr>
<td>11</td>
<td>153</td>
<td>had &lt;cons=a&lt;/cons&gt;</td>
</tr>
<tr>
<td>12</td>
<td>136</td>
<td>I have &lt;add&gt;</td>
</tr>
<tr>
<td>13</td>
<td>131</td>
<td>t have</td>
</tr>
<tr>
<td>14</td>
<td>130</td>
<td>msf have</td>
</tr>
<tr>
<td>15</td>
<td>109</td>
<td>don’t have</td>
</tr>
<tr>
<td>16</td>
<td>100</td>
<td>have &lt;add=a&lt;/add&gt; &lt;breakfast</td>
</tr>
<tr>
<td>17</td>
<td>99</td>
<td>msf have</td>
</tr>
<tr>
<td>18</td>
<td>96</td>
<td>n’t have</td>
</tr>
</tbody>
</table>

Total No. of Cluster Types: 14776, Total No. of Cluster Tokens: 29169
Errors related to ‘have’

- The n. of article additions (218) is almost the same as that of omissions (213):
  - “have a …” forms an unanalyzed chunk
  - “have *a breakfast”/ “have *a time to …”

- Also the negation errors are very frequent:
  T: So I don't have time to eat breakfast
  O: So I have n't time to eat breakfast
  A: So I <oms>don't</oms> have <add>n't</add> time to eat breakfast
Supervised vs. unsupervised learning

Automatic extraction of error patterns from LC

Multivariate Analysis

Supervised Learning
- Classification
  - Advanced
  - Intermediate
  - Novice

Unsupervised Learning
- Clustering
  - Errors?
New project: ICCI

- International Corpus of Crosslinguistic Interlanguage
- TUFS Global-COE Projects (5-year government-funded project)
- Aims: compiling corpora of young learners of English, comparable to JEFLL
- 7 countries (China; Taiwan; Israel; Spain; Poland; Austria; Singapore) at the moment
- Looking for more partner countries
ICCI: Comparable English learner corpora

JEFLI

- beginning – intermediate levels
- JH1 (year 7) – SH3 (year 12)
- 10,000 subjects; 670,000 words

JEFLI

- Spain, Austria, Israel, Poland, Taiwan, Hong Kong, Singapore

JEFLI

- Korea, China, Russia, France, etc.