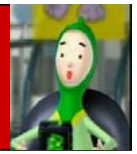


Error tagging

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

Error tagging: inline annotation

<A><F>Oh</F>. I see. Never been there. <F>Eh</F>, is that <g_at crr="a"></g_at> nice place?</A>

<B>Kind of.</B>

<A>Kind <OL>of</OL>.</A>

<B><OL><R>Ye</R></OL>, yes. It's <g_at crr="the"></g_at> <g_n_num crr="suburbs">suburb</g_n_num> .</B>

<A><OL>Suburb</OL>.</A>

<B><OL><??></??></OL>, <F>mm</F>. <?>But</?>, <g_at crr="a"></g_at> beautiful place.</B>

<A><F>Mhm</F>. I see. What did you <o_oms crr="do"></o_oms> there?</A>

<B>I studied. Yes, I <lxc crr="went to"><g_v_tns crr="entered">enter</g_v_tns></lxc> the university <o_oms crr="for"></o_oms> one year as <g_at crr="an"></g_at> exchange student.</B>

Error tagging schemes

- Depend on the research questions
- SST → surface grammatical/lexical errors
- Error tagging format:

```
<A>And, please describe this picture.</A>
<B>Describe? <F>Mhm</F>. <.></.> Maybe, <SC>this</SC>
<SC>toda</SC> <F>mm</F> it is a sunny day, and
<JP><F>unto</F></JP> in front of <SC?>hou</SC?>
<at odr="1" crr="a"></at> big house, <F>er</F> two
<n_inf odr="2" crr="housewives">housewifes</n_inf>
<v_agr odr="3" crr="are">is</v_agr> talking
<prp_lxc2 odr="4" crr="to"></prp_lxc2> each other.
```

<at odr="1" crr="a"> (article, order of correction, correct form)

Aspects of errors in the schemes

	CLC	**FreeText**	**Louvain**	*NICT JLE*
Phonology	–	–	–	–
Punctuation	√	√	√	–
Spelling	√	√	√	–
Grammar	√	√	√	√
Lexis	√	√	√	√
Pragmatics	√ (register)	√ (register)	√ (register)	√ (non-verbal cues)
Discourse	√ (pronoun reference)	√ (cohesion)	√ (unnatural discourse)	√ (self-corrections)

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,

FRIDA corpus tagging system

Error-tagged Learner Corpora and CALL: A Promising Synergy

Table 1
Error Domains and Categories

Error Domains		Error Categories	
<F>	Form	<AGL>	Agglutination
		<MAJ>	Upper/lower case
		<DIA>	Diacritics
		<HOM>	Homonymy
		<GRA>	Other spelling errors
<M>	Morphology	<MDP>	Derivation-prefixation
		<MDS>	Derivation-suffixation
		<MFL>	Inflection
		<MFC>	Inflection-confusion
		<MCO>	Compounding
<G>	Grammar	<CLA>	Class
		<AUX>	Auxiliary
		<GEN>	Gender
		<MOD>	Mode
		<NBR>	Number
		<PER>	Person
		<TPS>	Tense
		<VOI>	Voice
		<EUF>	Euphony
<L>	Lexis	<SIG>	Meaning
		<CPA>	Adjective complementation
		<CPD>	Adverb complementation
		<CPV>	Verb complementation
		<CPN>	Noun complementation
		<FIG>	Prefab
<X>	Syntax	<ORD>	Word order
		<MAN>	Word missing
		<RED>	Word redundant
		<COH>	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		

Granger (

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:

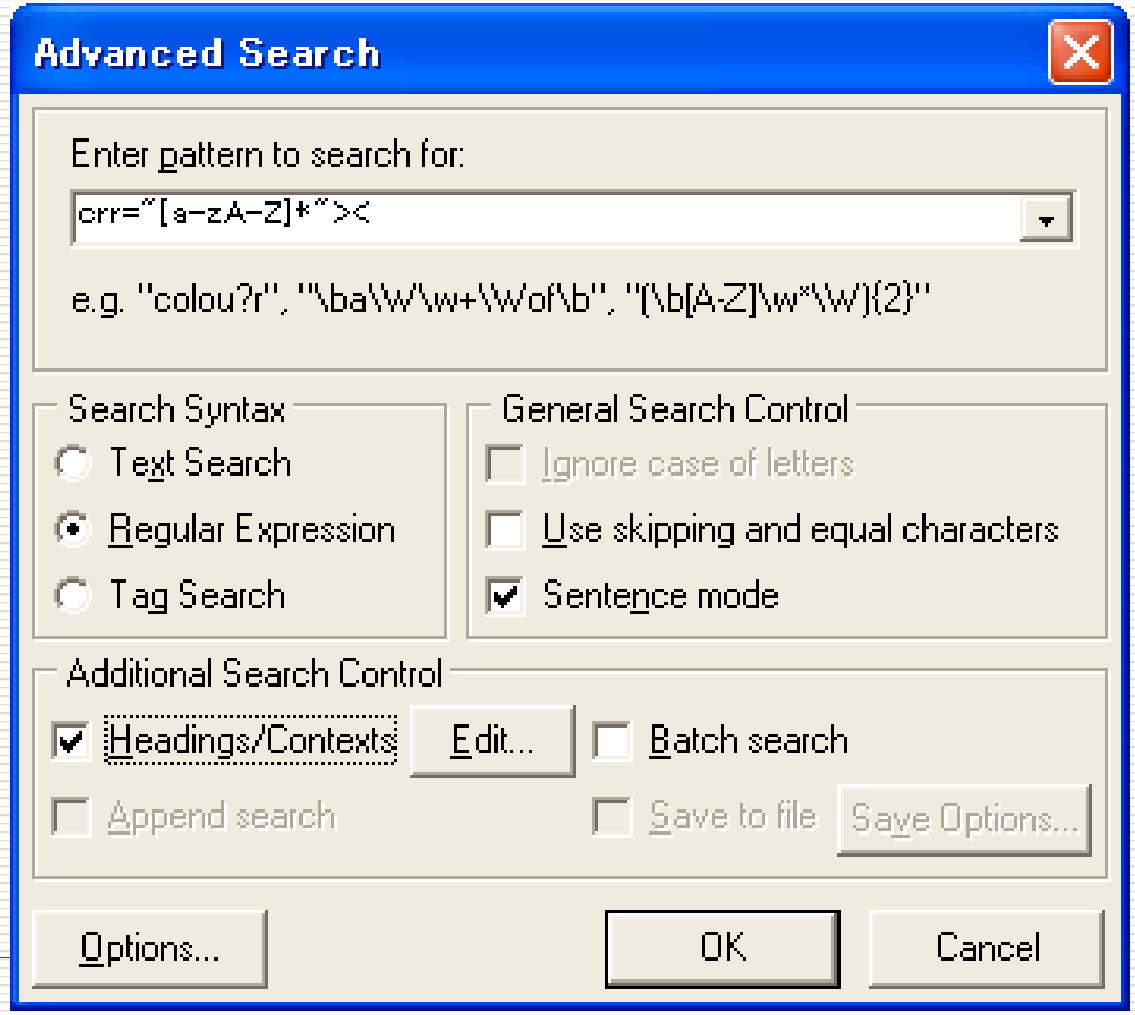
i) *pronoun*
ii) *conjunction*
iii) *determiner*
iv) *adjective*
v) *noun*
vi) *quantifier*
vii) *preposition*
viii) *verb*
ix) *adverb*.

The resulting annotation is shown in (9):

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

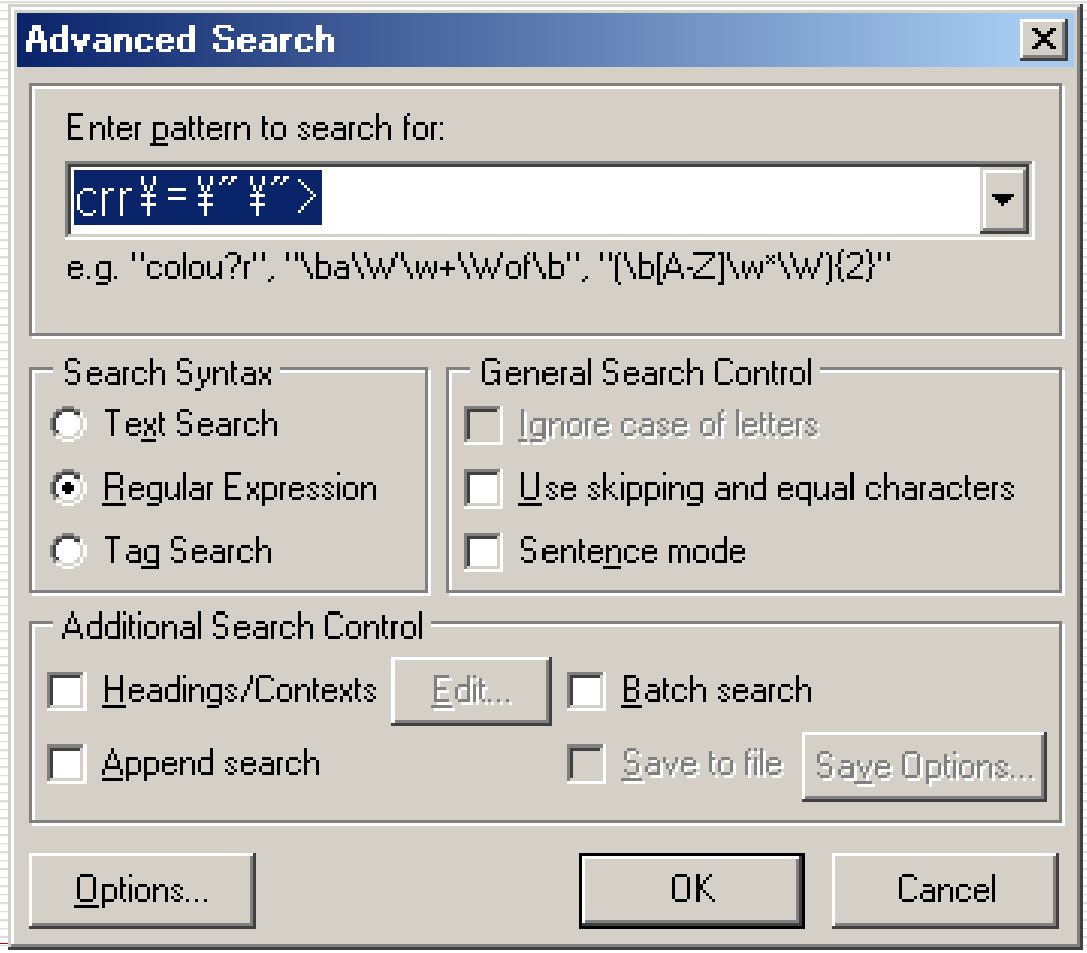
Searching for different error types

- 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

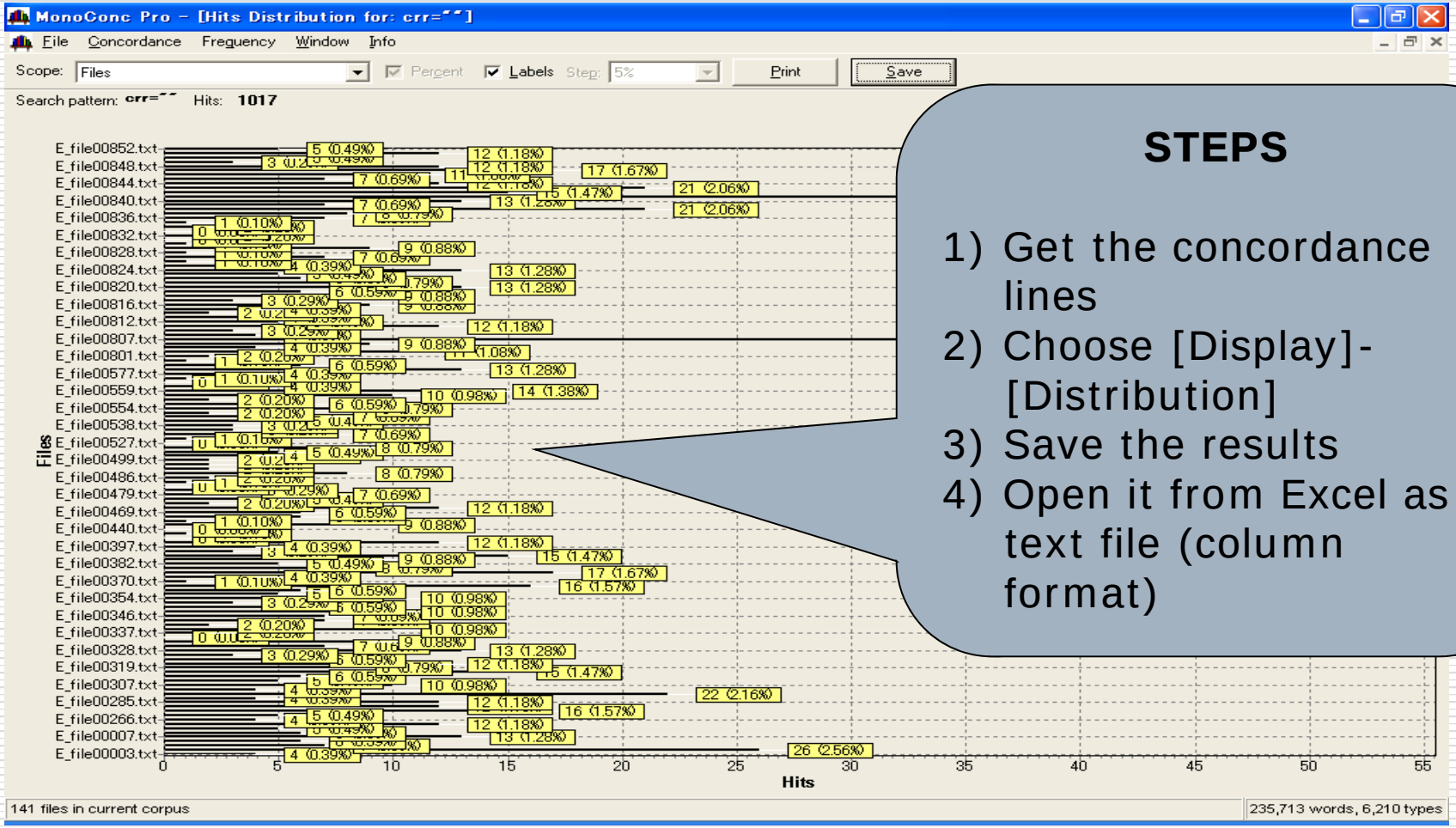


Searching for different error types

- **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

Microsoft Excel - addition_err_distrib...

B3 = 4

	A	B
1	File name	ts
2	----------	--
3	E_file00003.txt	4
4	E_file00004.txt	26
5	E_file00005.txt	7
6	E_file00006.txt	6
7	E_file00007.txt	13
8	E_file00008.txt	6
9	E_file00107.txt	5
10	E_file00127.txt	12
11	E_file00266.txt	4
12	E_file00269.txt	5
13	E_file00273.txt	16
14	E_file00276.txt	12
15	E_file00285.txt	12
16	E_file00286.txt	4
17	E_file00287.txt	22
18	E_file00301.txt	4
19	E_file00307.txt	10
20	E_file00309.txt	5
21	E_file00310.txt	6
22	E_file00318.txt	15
23	E_file00319.txt	8
24	E_file00320.txt	12
25	E_file00325.txt	6

addition_err_distribute

合計=214

Microsoft Excel - sst_level

C2 = 4

	A	B	C	D
1	file_name	SST_level	addition	
2	E_file00003.txt	5	4	
3	E_file00004.txt	4	26	
4	E_file00005.txt	4	7	
5	E_file00006.txt	6	6	
6	E_file00007.txt	6	13	
7	E_file00008.txt	6	6	
8	E_file00107.txt	6	5	
9	E_file00127.txt	8	12	
10	E_file00266.txt	4	4	
11	E_file00269.txt	4	5	
12	E_file00273.txt	7	16	
13	E_file00276.txt	4	12	
14	E_file00285.txt	7	12	
15	E_file00286.txt	7	4	
16	E_file00287.txt	9	22	
17	E_file00301.txt	7	4	
18	E_file00307.txt	6	10	
19	E_file00309.txt	4	5	
20	E_file00310.txt	4	6	
21	E_file00318.txt	7	15	
22	E_file00319.txt	9	8	
23	E_file00320.txt	8	12	
24	E_file00325.txt	9	6	
25	E_file00327.txt	7	3	
26	E_file00328.txt	6	13	
27	E_file00329.txt	6	7	
28	E_file00330.txt	7	9	

sst_level

コマンド　合計=252

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

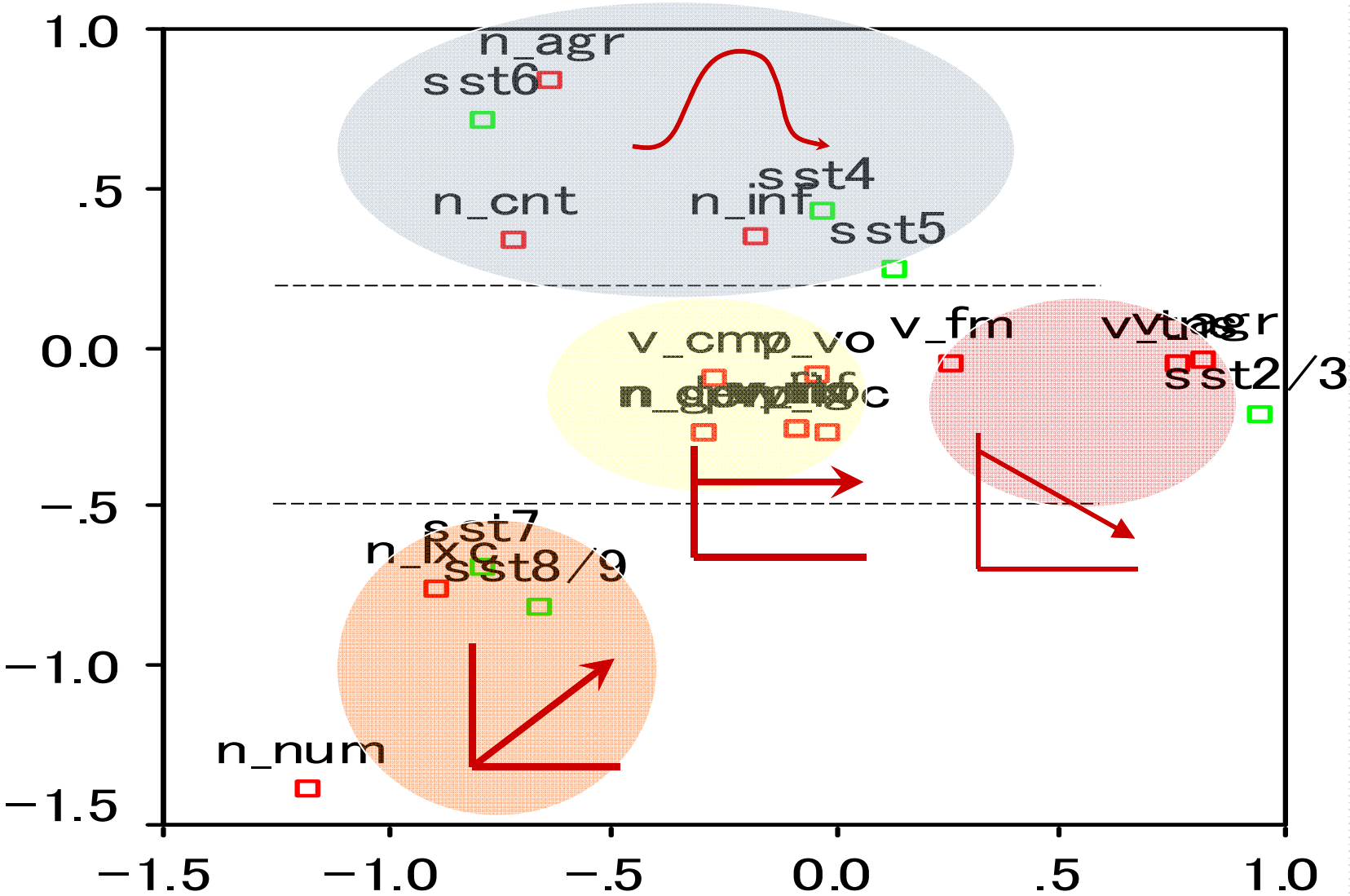
- Overuse vs. underuse

- 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

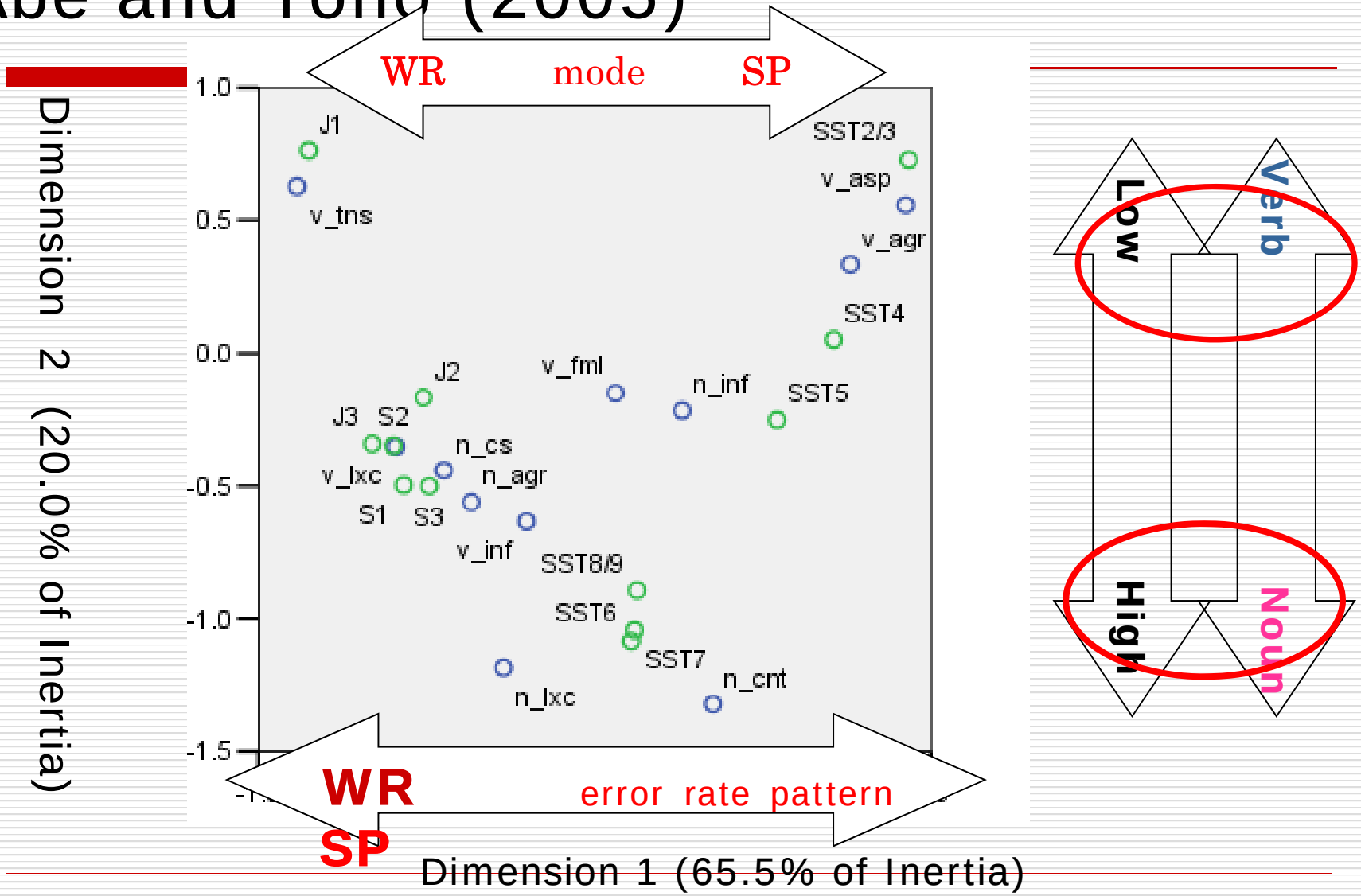
- Description → Explanation
- SLA theories:
  - UG-Based SLA (Hawkins, White) ← more focus on lexicon
  - Processibility Hypothesis (Pienemann) ← Levelt & LFG
  - Competition Model (MacWhinney) ← very much frequency-based
- Related disciplines:
  - Cognitive linguistics; Usage-based approach
  - Systemic-functional grammar
  - Natural language processing
  - Data mining; Neural network

Error freq's & distributions

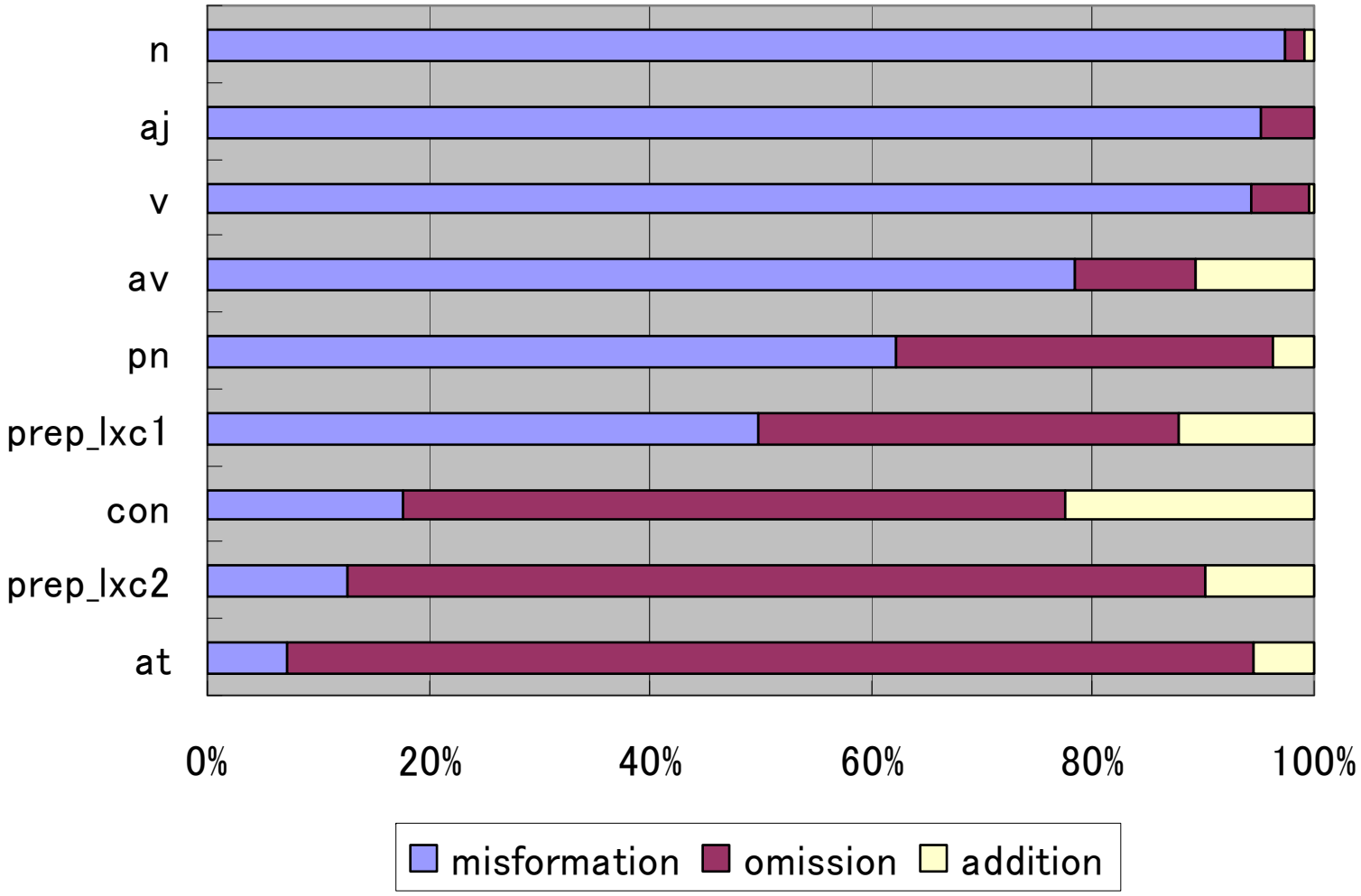


Dimension 1（64．2%） Dimension 2（寄与率20．6%）

Abe and Tono (2005)



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

ParaConc - [Parallel Concordance - [is]]

File Search Frequency Display Sort Window Info

... /text> <text> Our school festival day is Stember <jp> 10 gatu </jp> the 23, 24 ...
... . </text> <text> Our school festival is october 23 and 24. It festival has a c ...
... . </text> <text> Our school festival is October 23 and October 24. <jp> 24 nit ...
... . </text> <text> Our school festival is October 23 and 24. Every class sings o ...
... /text> <text> Our school art festival is Friday October 23 and Satuday Octover 2 ...
... <jp> no bijutu seikaku </jp>. Festival is October 23 and 24. What is the date th ...
... ting. </text> <text> School festival is October 23 and October 24. Every class ...
... . </text> <text> Our school festival is October 23, 24 Every class sings one s ...
... > </text> <text> Our school festival is October 23rd and 24th. Our school fest ...
... hool festival has a chorus contest. It is in Friday and Saturday. It is <jp> med ...
... song contert is fastval mein. Festival is octover 23 and 24 in my school. My cla ...
... hen. </text> <text> Our school music is October the tenty tree. My class song ...
... > </text> <text> Our school festival is Fryday October 23 and 24. Our class ma ...
... . </text> <text> Our school festival is October 23 and 24. We are sing a song ...
... ! </text> <text> Our school festival is Octorber 23 and 24. 24 is <jp> medama ...
... . </text> <text> Our school festival is October 23 and October 24 <jp> ni yarim ...
... ober 23 and October 24. chorus contest is October 24. Iam a member of a music cl ...
... tame <jp> (teema) </jp> song. Festival is October 23 and October 24. chorus cont ...
... ! </text> <text> Our school festival is October 23 and 24. Our class sing <jp> ...
... ! </text> <text> Our school festival is October 23 and 24. It is fun very much ...
... ext> <text> My school festibal. Date is Friday October 23 and Saturday October ...

ong was "kimi wo nosete". </text> <text> Our school festival was on October 23rd and 24th. It had a ...
... a very very happy day. </text> <text> Our school festival was on October 23rd and October 24th. On the ...
musical. It was very nice. </text> <text> Our school festival was on October 23rd and 24th. Every class sang ...
al was very interesting. </text> <text> Our school art festival was on Friday, October 23 and Satuday, October 24. ...
... song contest and a picture competition. The festival was on October 23rd and 24th. What is the ...
ntest. But I was interested. </text> <text> The school festival was on October 23rd and October 24th. Every class ...
m, there was an exhibition. </text> <text> Our school festival was on October 23rd and 24th. Every class sang ...
onomics and artworks. (?) </text> <text> Our school festival was on October 23rd and 24th. Our school festival ...
... Our school festival had a song contest. It was on Friday and Saturday. It was the highlight ...
... contest were the festival's main events. The festival was on October 23rd and 24th in my school. ...
. songs at that time. </text> <text> Our school music festival was on October the twenty third. My class' song ...
wasn't completely satisfied. </text> <text> Our school festival was on Friday, October 23rd and the 24th. Our ...
port. Its theme was herbs. </text> <text> Our school festival was on October 23rd and 24th. We sang a ...
joyed the school festival!! </text> <text> Our school festival was on October 23rd and 24th. The 24th was ...
puldn't stand on the stage. </text> <text> Our school festival was on October 23rd and October 24th. In our ...
... October 23rd and October 24th. The song contest was on October 24th. I am a member of ...
... It's the theme from "Tenkunohoshi Rapyuta". The festival was on October 23rd and October 24th. The song ...
enjoying myself very much. </text> <text> Our school festival was on October 23rd and 24th. Our class sang ...
school festival very much! </text> <text> Our school festival was on October 23rd and 24th. It was really ...
flight. It was very beautiful. </text> <text> My school festival was on Friday, October 23rd and Saturday, October 24th. ...

500 matches | English (United States) - Search word, 1st right | Strings matching: is

English (United States) - C:¥corpora¥jefll.scn.errortagged¥parallel¥001-500.corr.txt, line: 2008 | 16,297 / 17,851

Upper pain: original vs. lower pain: corrected

Errors involved in copula "*be*"

ParaConc - [Parallel Concordance - [is]]

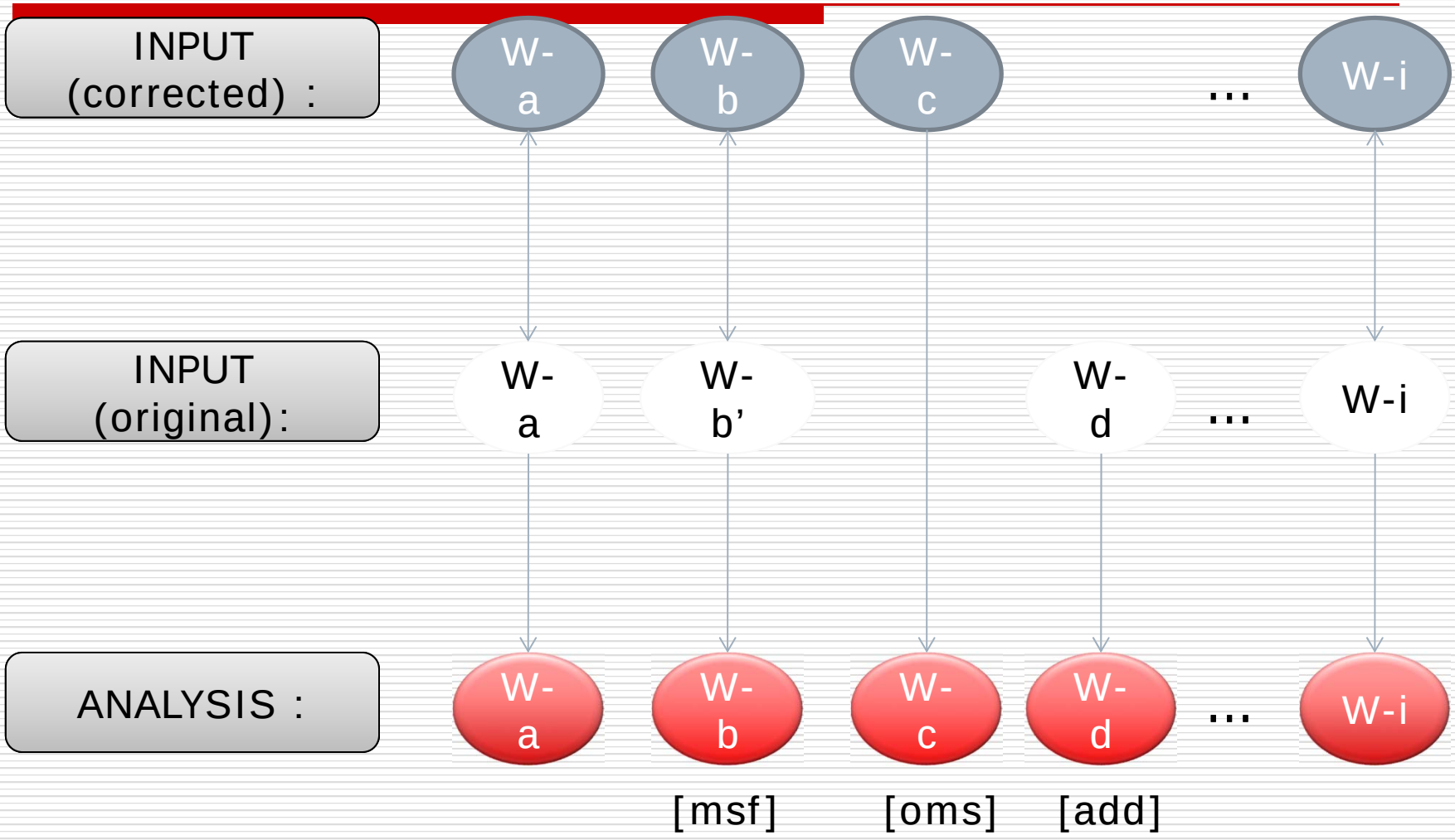
File Search Frequency Display Sort Window Info

... iday, october 23 and Satday, october 24 is my school festival day. It's very big ...
... very fun good <jp> omoide </jp>. club is <jp> butai happyou </jp>. It is very f ...
... hard. Our class had a <jp> tenji </jp> is <jp> mozaiku ga </jp>. very <jp> taihe ...
... imasita </jp>. <jp> gasshoukyoku </jp> is <jp> "kimi wo nosete" </jp>. class <jp ...
... jp> ni okonawaremasita </jp>. My class is <jp> gasshou </jp> and classarts <jp> w ...
... club is musical are very well. I enjoy is it festival! </text> <text> Our scho ...
... it at school every day. 1C sings songs is "KIMIONOSETE". It's very butiful song. ...
... /jp>. very very enjoy day my up class is <jp> rittai sakuhin </jp>. it is big ...
... wo happyou simasita </jp>. It's lesson is very besy. But chorus club's <jp> happ ...
... > My school festival is veri nice. It is <jp> tenji sita </jp> big art our class ...
... October 23 and 24. We are sing a song is <jp> "Kimi wo nosete" </jp> It is very ...
... ut I don't like <jp> pikuchaa </jp>. I is don't see a little <jp> "miniku katta" ...
... tuhin </jp> contest. <jp> teema </jp> is <jp> habataki </jp> <jp> kakumei </jp> ...
... > dasimono </jp> in <jp> habataki </jp> is play. <jp> medama </jp> of the fstiva ...
... . </text> <text> Our school festival is <jp> harie </jp> and chorus contest. c ...
... kai </jp>. <jp> happyou sita </jp> day is October the twenty third, and twenty fo ...
... ext> <text> Our school festival Today is October 23 24. It is our <jp> doryoku ...
... Every class sings one songs. My class is <jp> "Kimi wo nosete" </jp>. I like it ...
... happyou </jp>. But 1-A, 1-B, 1-C, 1-D is not <jp> kekka happyou </jp>. We were ...
... Every class sings one songs. Many club is <jp> happyou </jp>. Our class had a <j ...
... stival is look at <jp> okyaku san </jp> is very very <jp> ooi </jp>. <jp> sono ko ...

Friday, October 23rd and Saturday, October 24th were my school festival days.
I am in the on-stage performance club. (?)
Our class had a mosaic exhibition.
We sang "kimi wo nosete" in the song contest.
My class took part in the song and art contests. (?)
I enjoyed the festival!
1C sang the song "Kiminosete".
My class made a 3-D artwork. (?)
We were very busy doing rehersals for it. (?)
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 the 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.

500 matches | English (United States) - Search word, 1st right | Strings matching: is
English (United States) - C:¥corpora¥jefll.scn.errortagged¥parallel¥001-500.corr.txt | 16,297 / 17,851

DP matching


INPUT (corrected) :
W-a
W-b
W-c
...
W-i
INPUT (original):
W-a
W-b'
W-d
...
W-i
ANALYSIS :
W-a
W-b
W-c
W-d
...
W-i
[msf]
[oms]
[add]

Automatic identification of learner errors

The first reason is every member of my family is busy in t

first reason is every my family is busy in the morning

<oms>The</oms> <oms>member</oms> <oms>of</oms>

- Looking at n-grams for maximum match and analyse t unmatched elements:

Automatic identification: output

T: My mother cooks very well ← corrected sentence

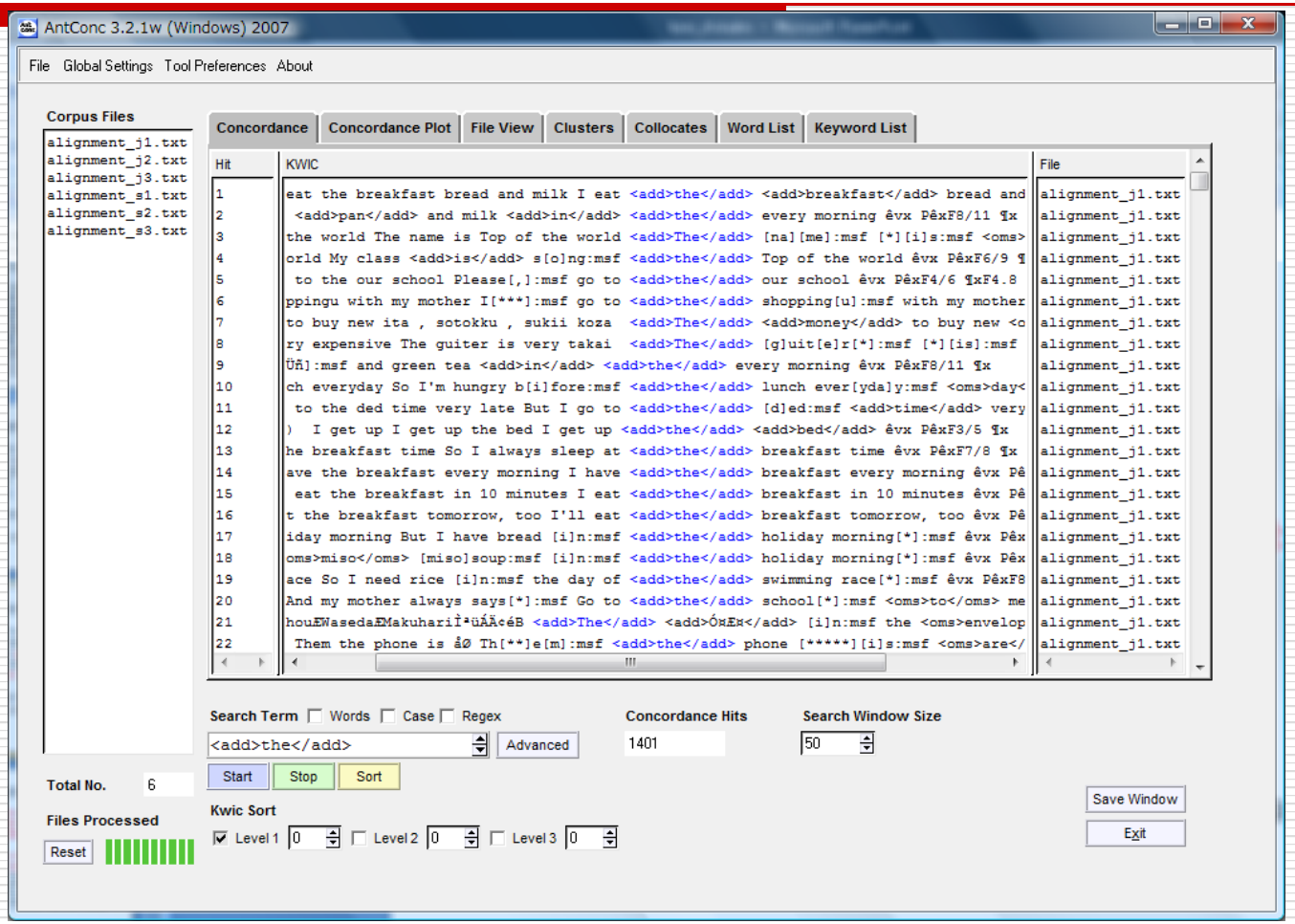
O: mother is cook very well ← original sentence

A: <oms>My</oms> mother <add>is</add> cook[*]:msf very well ← identifying differences

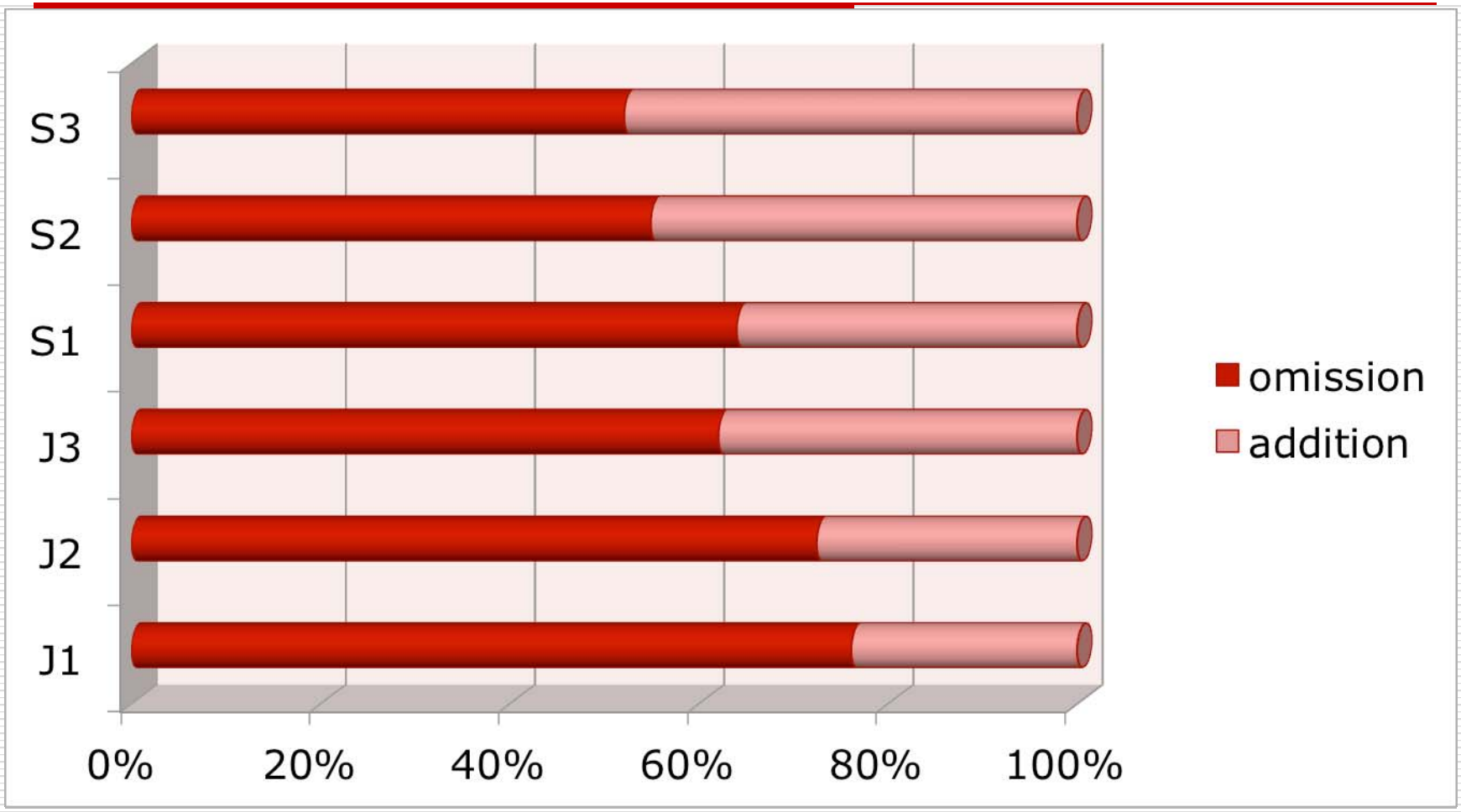
- Correspondence ratio:
  - Word level: 3/5
  - Character level: 3.80/5(76%)

Notes: T = target; O = original; A = analysis

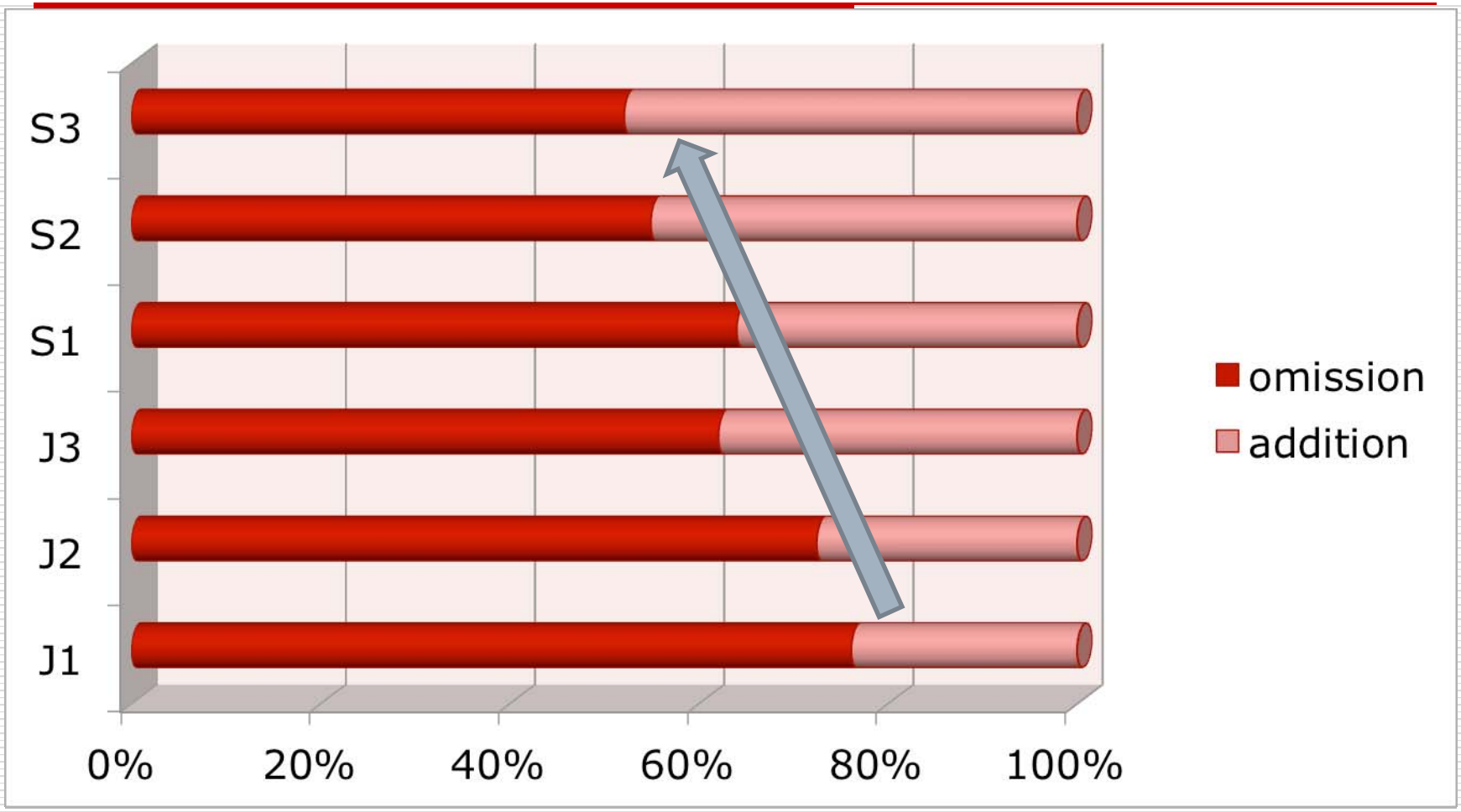
Looking for criterial features



Distributions of error types

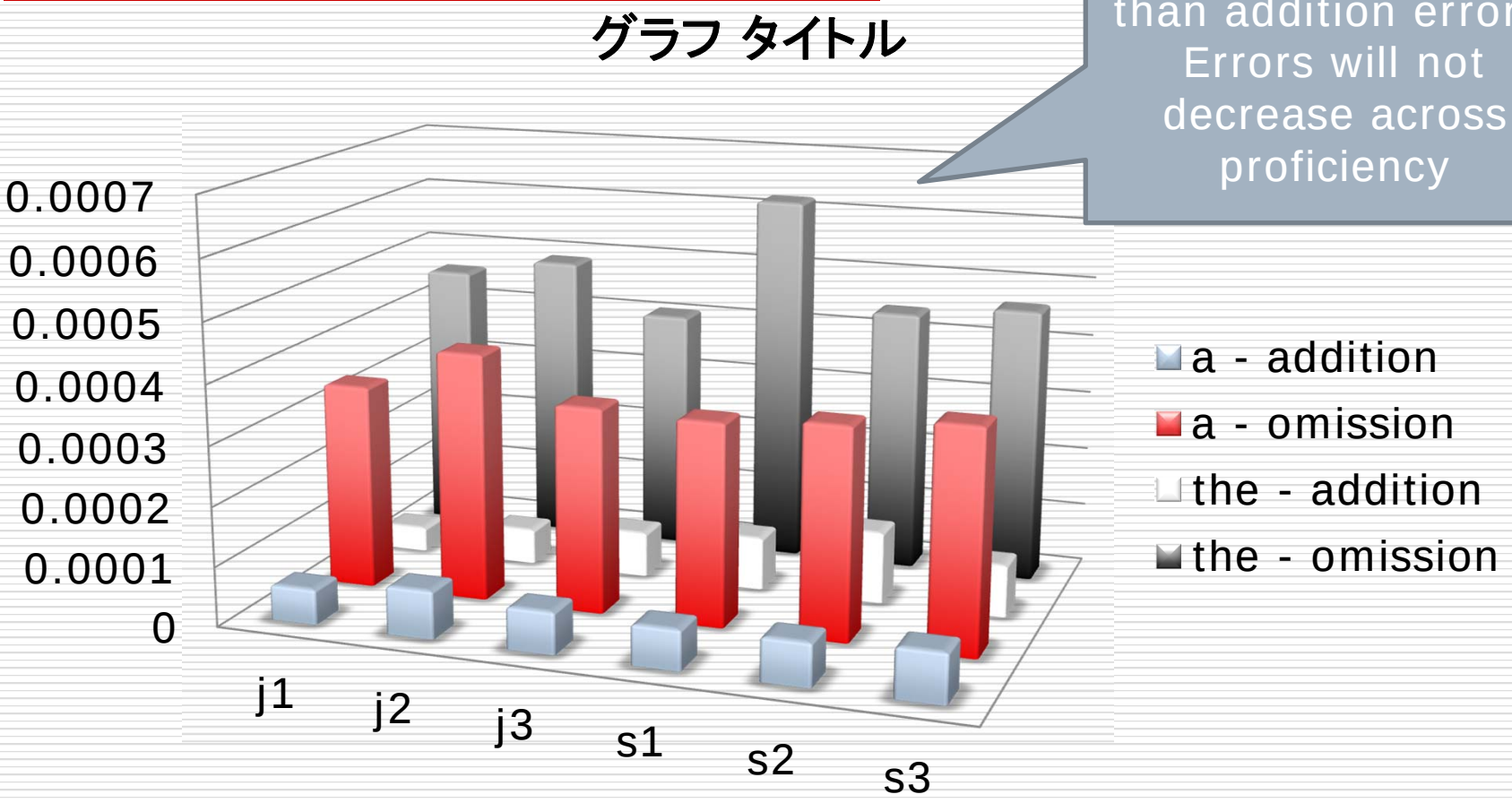


Distributions of error types



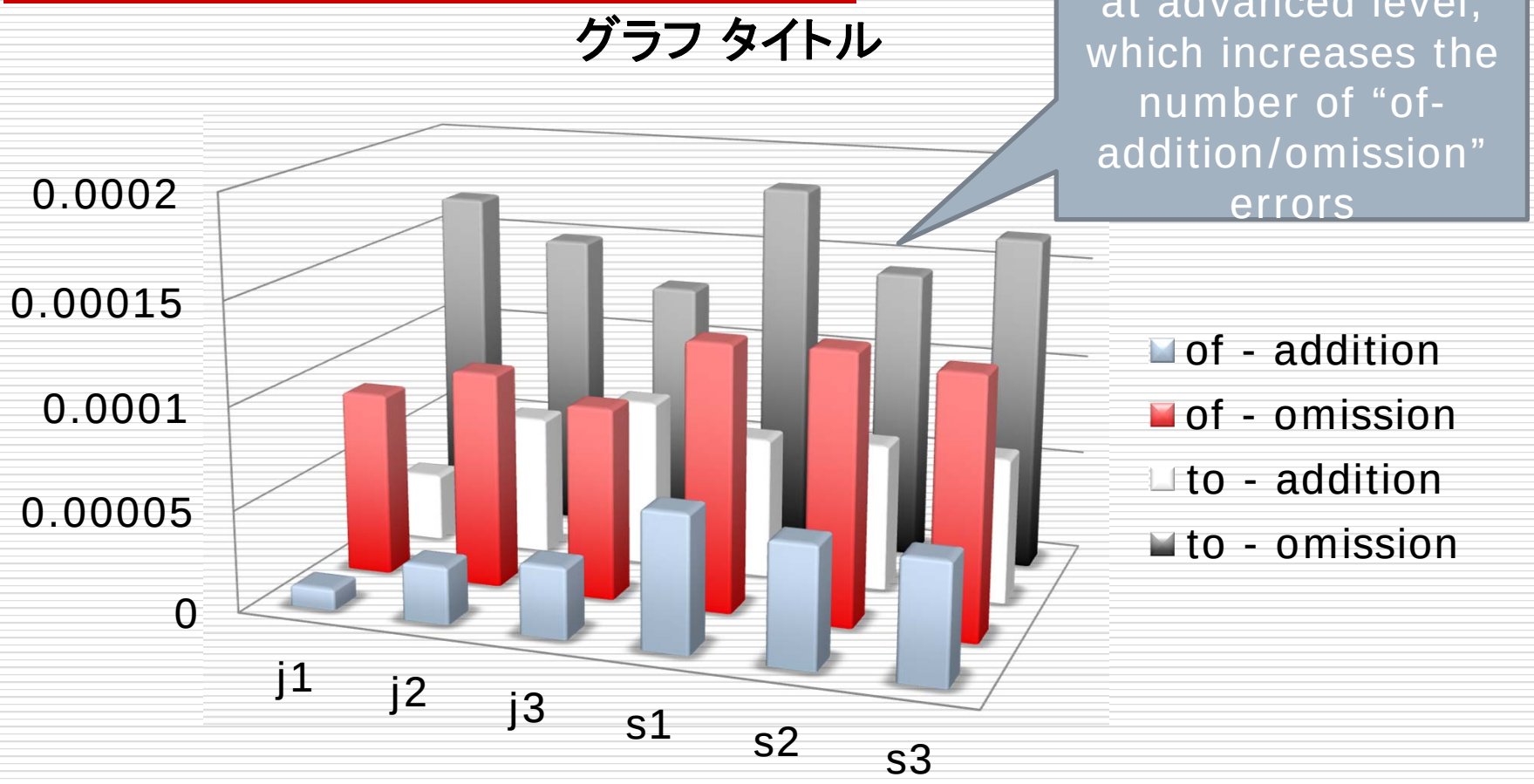
Article errors

Omission errors are far more common than addition errors. Errors will not decrease across proficiency



Omission errors are significantly more frequent than addition errors.

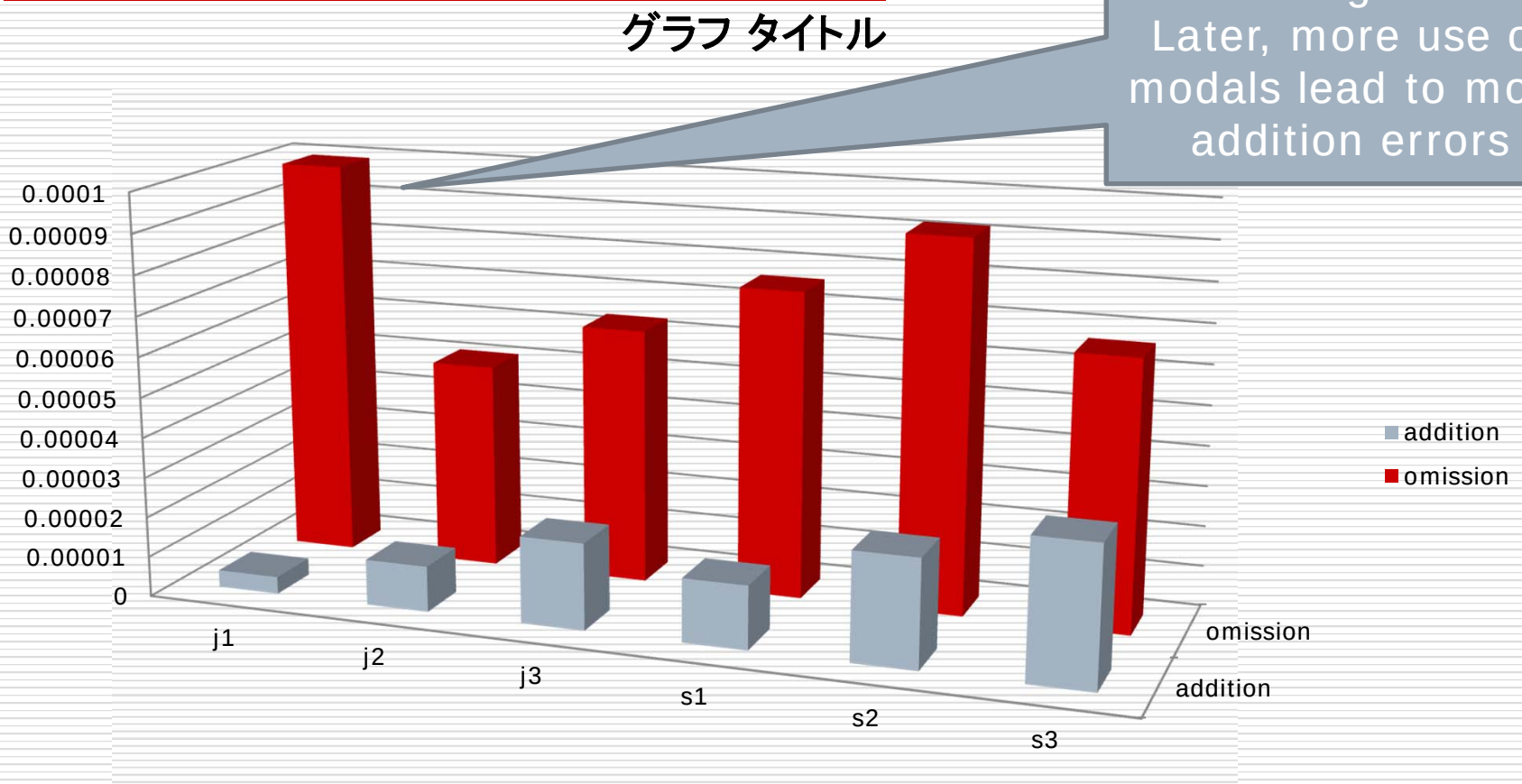
Preposition errors (to/of)



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’

AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

alignment_j1.txt
alignment_j2.txt
alignment_j3.txt
alignment_s1.txt
alignment_s2.txt
alignment_s3.txt

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Total No. of Cluster Types: 14776 Total No. of Cluster Tokens: 28169

Rank	Freq	Cluster
1	515	have <oms
2	388	have <add
3	303	I have
4	245	had <oms
5	218	have <add>a
6	218	have <add>a</add
7	213	have <oms>a
8	213	have <oms>a</oms
9	167	I have <oms
10	153	had <oms>a
11	153	had <oms>a</oms
12	134	I have <add
13	131	t have
14	130	msf have
15	109	don't have
16	100	have <add>a</add> breakfast
17	99	t:msf have
18	96	n't:msf have

N-word clusters of “have”

Search Term Words Case Regex N-Grams

(have|has|had|having) \<

Advanced

Start Stop Sort

Sort by

Sort by Freq

Search Term Position

On Left On Right Invert Order

Cluster Size

Min. Size 2 Max. Size 7

Min. Cluster Frequency

1

Save Window

Exit

Total No. 6

Files Processed

Reset

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

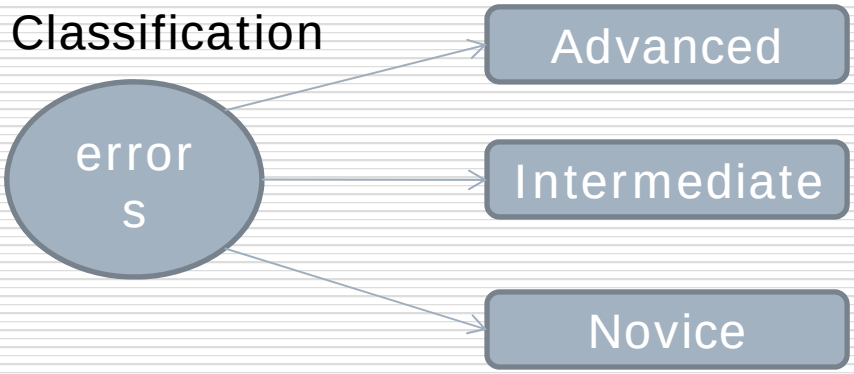
Automatic extraction of error patterns from LC

Multivariate Analysis

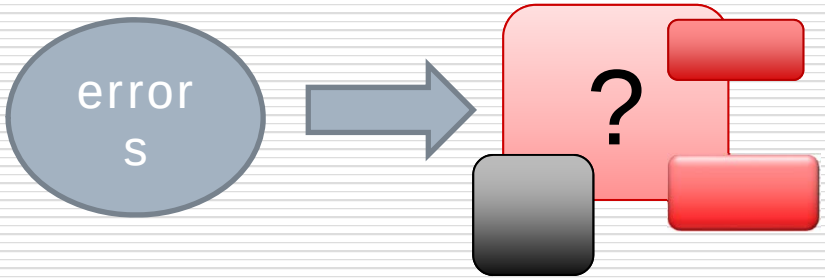
Supervised Learning

Unsupervised Learning

Classification



Clustering



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

JEFLL

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