Corpus Linguistics (4): Learner corpora & error tagging

Yukio Tono (TUFS)

LCR and related fields

Corpus linguistics	• What is a corpus? • How can we use a corpus?
SLA	• How can we make sense of the data? • What answer are we looking for?
ELT	• How can we apply corpus findings? • How can we improve our teaching?

Three corpus-linguistic methods

1. Frequency lists & collocate lists
   - Most decontextualized methods
2. Colligations (Collostructions)
   - Lexical elements + grammatical element or structure
3. Concordances (of search expressions)
   - The occurrence of a match of the search expression
   - Most context-rich

Important concepts in CL (2)

- Frequency vs distribution (dispersion)
- Collocation (lexical n-grams; prefabs; multi-word units)
  - node vs collocate
  - N-word cluster
- Colligation
  - Lexico-grammatical co-occurrence
- Concordance
  - KWIC; Search [node] –word; right/left contexts; sorting;

Data relevant to SLA/ELT

- How does the input language pattern?
- How does the native language of the learners pattern?
- How does the target language pattern?
- What are the differences between how the native language and the target language pattern? (Contrastive/Cross-linguistic Analysis)

Data relevant to SLA/ELT

- How does the input language pattern?
  - Textbook corpus/ Classroom interaction corpus
- How does the native language of the learners pattern?
  - L1 corpus
- How does the target language pattern?
  - TL corpus (e.g. English native corpus)
- What are the differences between how the native language and the target language pattern? (Contrastive/Cross-linguistic Analysis)

Important concepts in CL (3)

- General vs. Specialized corpora
- Spoken vs. Written corpora
- Balanced vs. Monitor corpora

Output of the learner

- How does the interlanguage pattern?
  - Learner corpora
- Which kinds of errors do the language learners commit?
  - Computer-aided error analysis (Granger)

Types of Learner Corpora

- Proficiency levels:
  - Fixed vs. varied (cross-sectional/longitudinal)
- L1 background:
  - Fixed vs. varied (learners with various L1s)
- Mode of production:
  - Written (essay)
  - Spoken (speech; retelling; conversation)
- Levels of annotation (POS; parsed; error-tagged)

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

LCR & ELT applications

- Indirect use:
  - Lexicography
  - Wordlist
  - Syllabus/course design
  - Materials design (textbooks; vocabulary books; classroom tasks)
  - Test development (CEFR; Criterion; SST)
- Direct use:
  - Corpus use in the classroom
  - Data driven learning
  - CALL implementations
  - Teacher training

Main areas to be covered

- State-of-the-art articles in LCR
- Error annotation
- CEFR-based LCR
- Automatic detection of errors using LC
- Applications of LCR in iCALL
- Spoken vs. Written LC

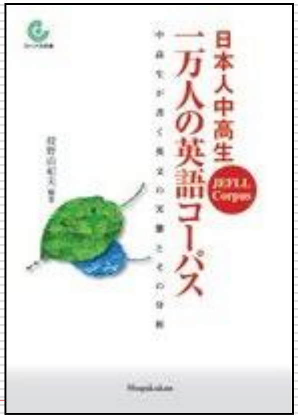
Error tagging

- Annotation on language learners' errors
- Error-tagged corpora:
  - NICT JLE Corpus: partially error-tagged
  - JEFLL Corpus: partially error-tagged
  - Cambridge Learner Corpus
  - HKUST Corpus of Learner English
- Generic error tagsets:
  - NICT JLE/ ICLE
- Tagging is usually done manually

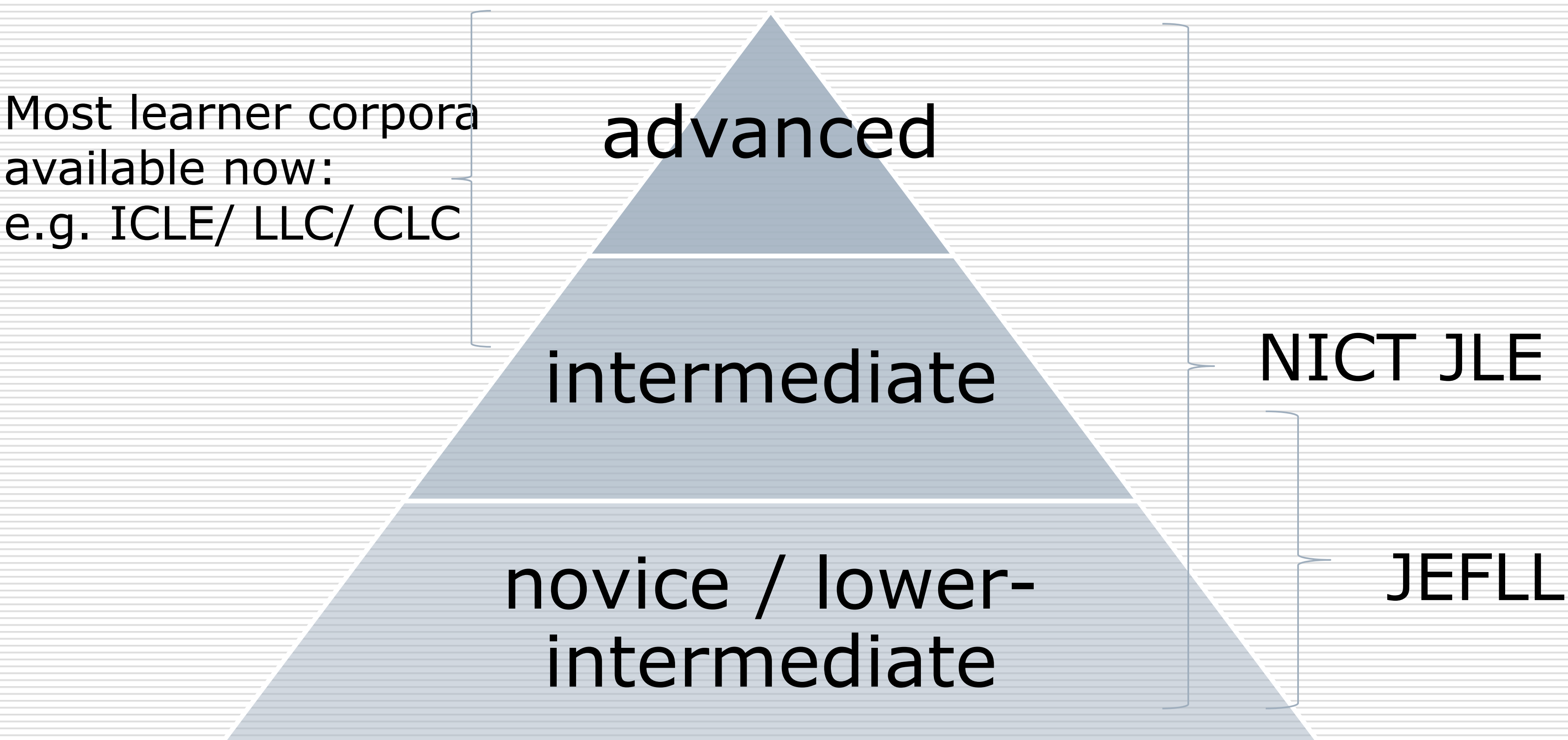
Learner corpus projects in Japan

- NICT JLE Corpus (Izumi et al. 2005)
  - 2 million words
  - Spoken
    (based on the OPI-like interview scripts)
  - 1,283 subjects
  - Distributed by NICT

- JEFLL Corpus (Tono et al. 2007)
  - 669,281 words
  - Written in-class essays (w/o dictionary)
  - 10,038 subjects (junior & senior high)
  - Freely accessible on the web:
  - http://scn02.corpora.jp/~jefll04dev/



L2 vocabulary profile: Crucial differences

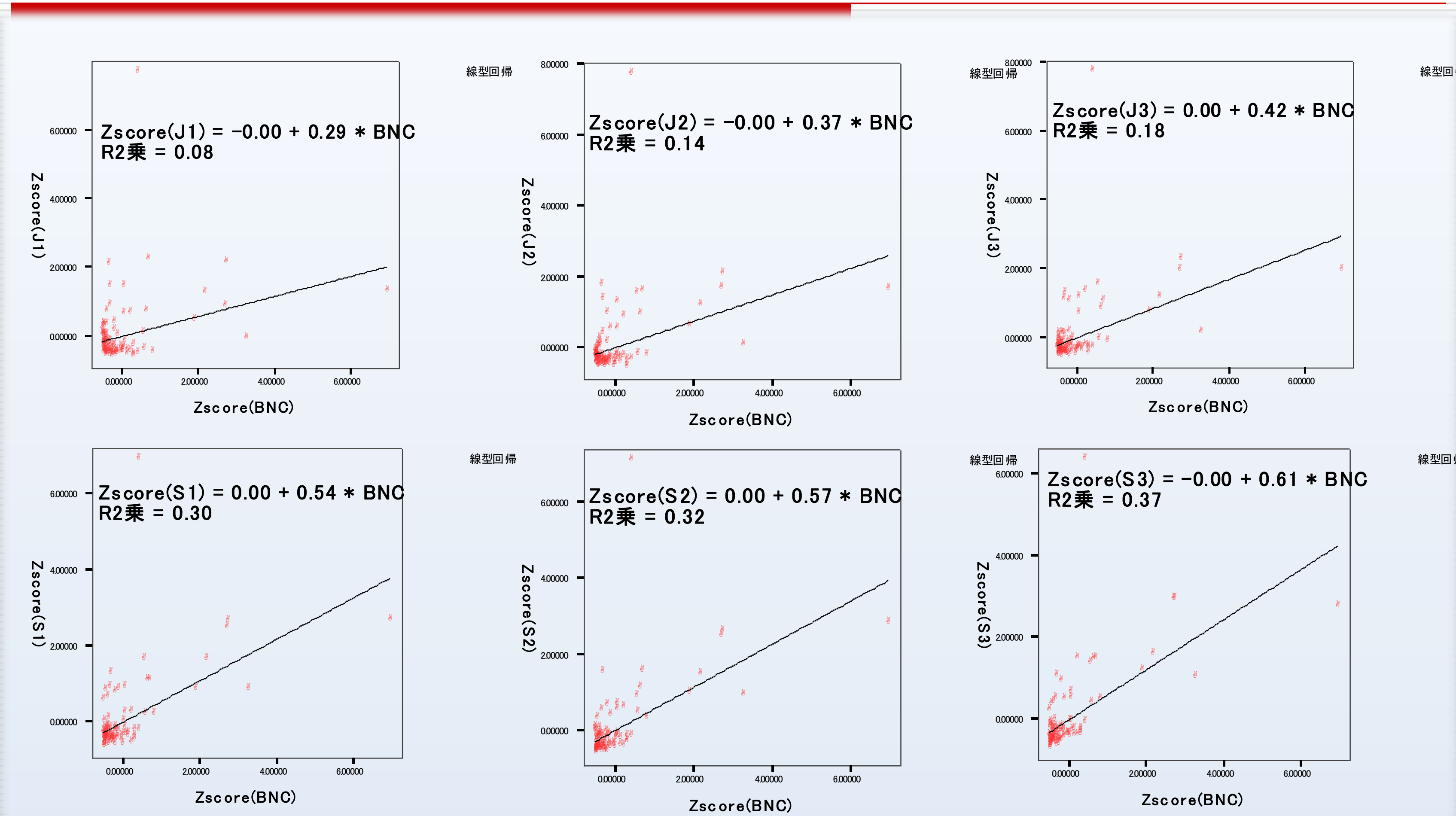


Systematizing LC descriptions

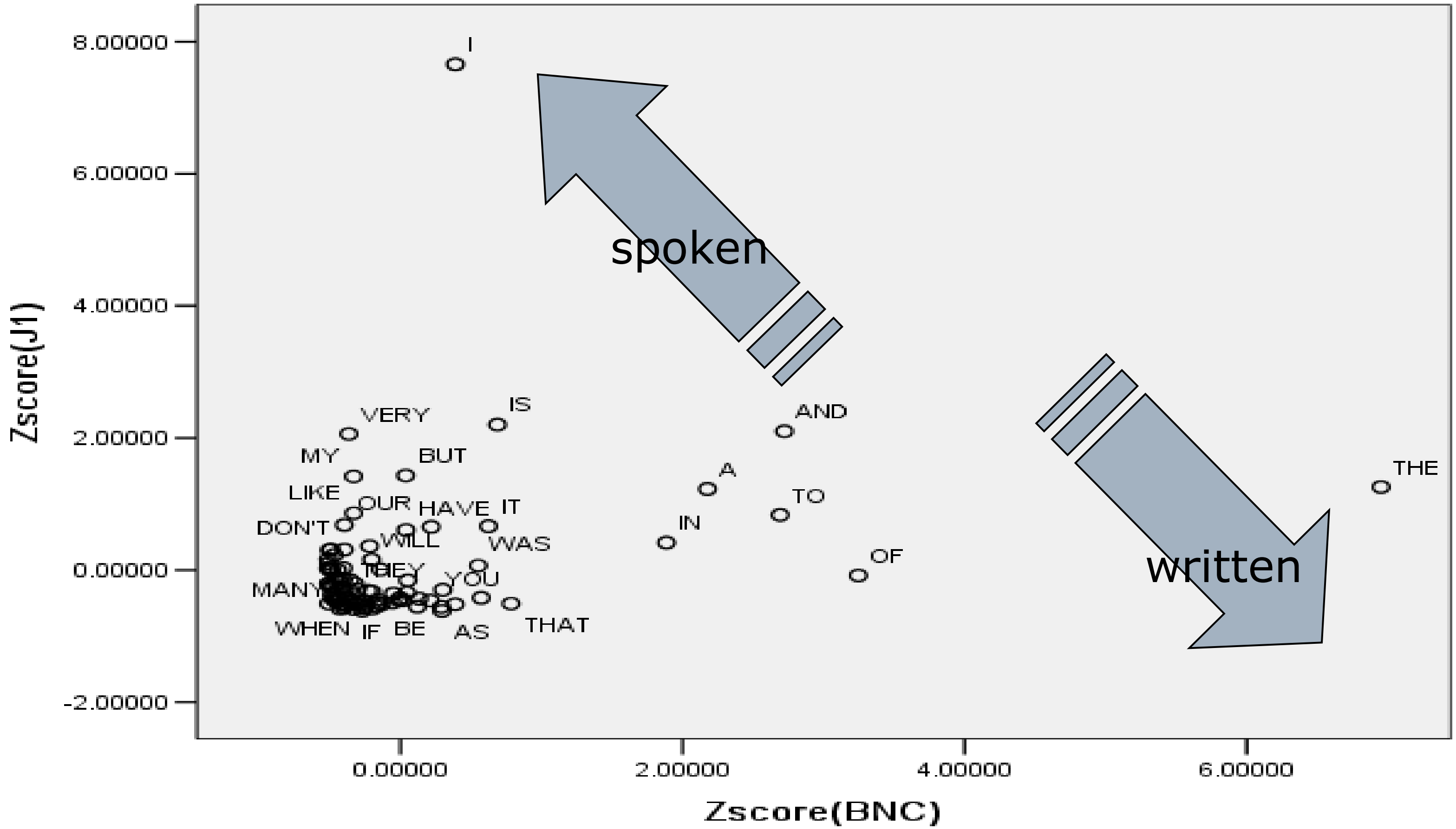
- A series of studies on criterial features of L2 developmental stages based on JEFLL and NICT JLE:
  - Morpheme orders: Tono (1998); Izumi (2005)
  - N-gram analysis: Tono (2000, 2008); Kimura (2004)
  - Verb subcategorization: Tono (2004)
  - Verb & noun errors: Abe (2003, 2004, 2005)
  - Article errors: Izumi (2003, 2004)
  - NP complexity: Kaneko (2004, 2006); Miura (2008)
  - Conjunctions: Kobayashi & Yamada (2008)

WORDLIST ANALYSIS

Use of top 100 words (JEFLL)



Distributions of top 100 words (JEFLL)



Overuse/underuse of words in top 10

J1		J2		J3		S1		S2		S3		BNC	
Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
I	7071	I	7235	I	7513	I	6106	I	6386	I	5190	THE	6178
IS	2412	AND	2350	AND	2624	THE	2706	THE	2899	AND	2726	OF	3116
AND	2328	VERY	2101	THE	2350	AND	2699	AND	[illegible]	TO	2722	AND	2682
VERY	2293	TO	2020	TO	2344	TO	[illegible]	TO	2618	THE	2580	TO	2656
BUT	1755	THE	1979	WAS	[illegible]	[illegible]	1898	IS	1850	A	1744	A	2229
MY	1743	IS	[illegible]	[illegible]	1791	WAS	1893	MY	1830	IS	1654	IN	1989
THE	1606	[illegible]	1867	MY	1730	MY	1597	A	1790	HE	1652	THAT	1075
A	1581	MY	1739	A	1641	IT	1460	IT	1500	IT	1648	IS	996
LIKE	1268	BUT	1666	BUT	1627	IS	1448	IN	1376	WAS	1574	IT	943
TO	1245	A	1573	VERY	1585	VERY	1316	OF	1334	IN	1451	FOR	900

freq =per 100,000 words

COLLOCATION/COLLIGATION ANALYSIS

The use of “make + Noun”

L2 vocabulary profile:LC perspectives

Structural complexities

- have + P.P.
- have + NP
- have + NP + V
- have + NP + Ving

HIGH

LOW

Semantic complexities

have + NP:

abstract N
delexical use

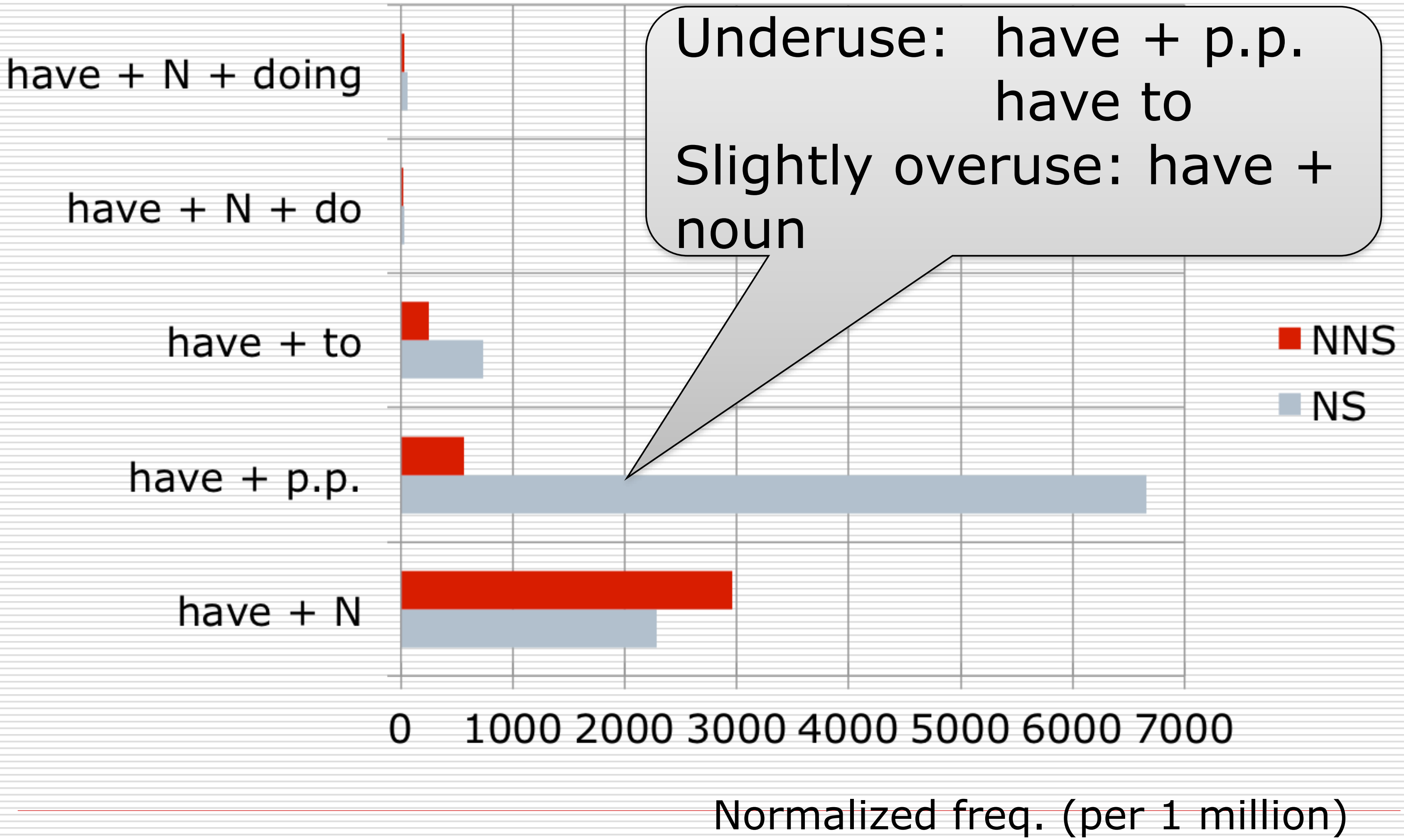
concrete N

Learner corpora

LC descriptions needed to describe the gap between NS and NNS performances

Profiling based on NS corpora only

Use of the verb *have*



have + noun

BNC Top 10	NS	JEFLL-JH	JEFLL-SH
have + **time**	*****	************	***************
have + **right(s)**	*****	-	-
have + **problem**	****	-	-
have + **effect**	****	-	-
have + **look**	****	-	-
have + **child**	***	-	-
have + **idea**	***	*	*
have + **chance**	***	-	-
have + **place**	***	-	-
have + **power**	***	-	-

have + noun (2)

JEFLL Top 10	NS	JEFLL-JH	JEFLL-SH
have + **breakfast**	*	*********************	*****************
have + **bread**	-	******************	****************
have + **rice**	-	*****************	**************
have + **time**	*****	************	***************
have + **money**	**	*********	*********
have + **dream**	-	*******	******
have + **food**	-	****	****
have + **lunch**	*	****	***
have + **break**	*	**	**
have + **thing**	*	**	*

have + noun (3)

Very little use of delexical verbs

(e.g. have a look)

Fewer abstract nouns

time

idea

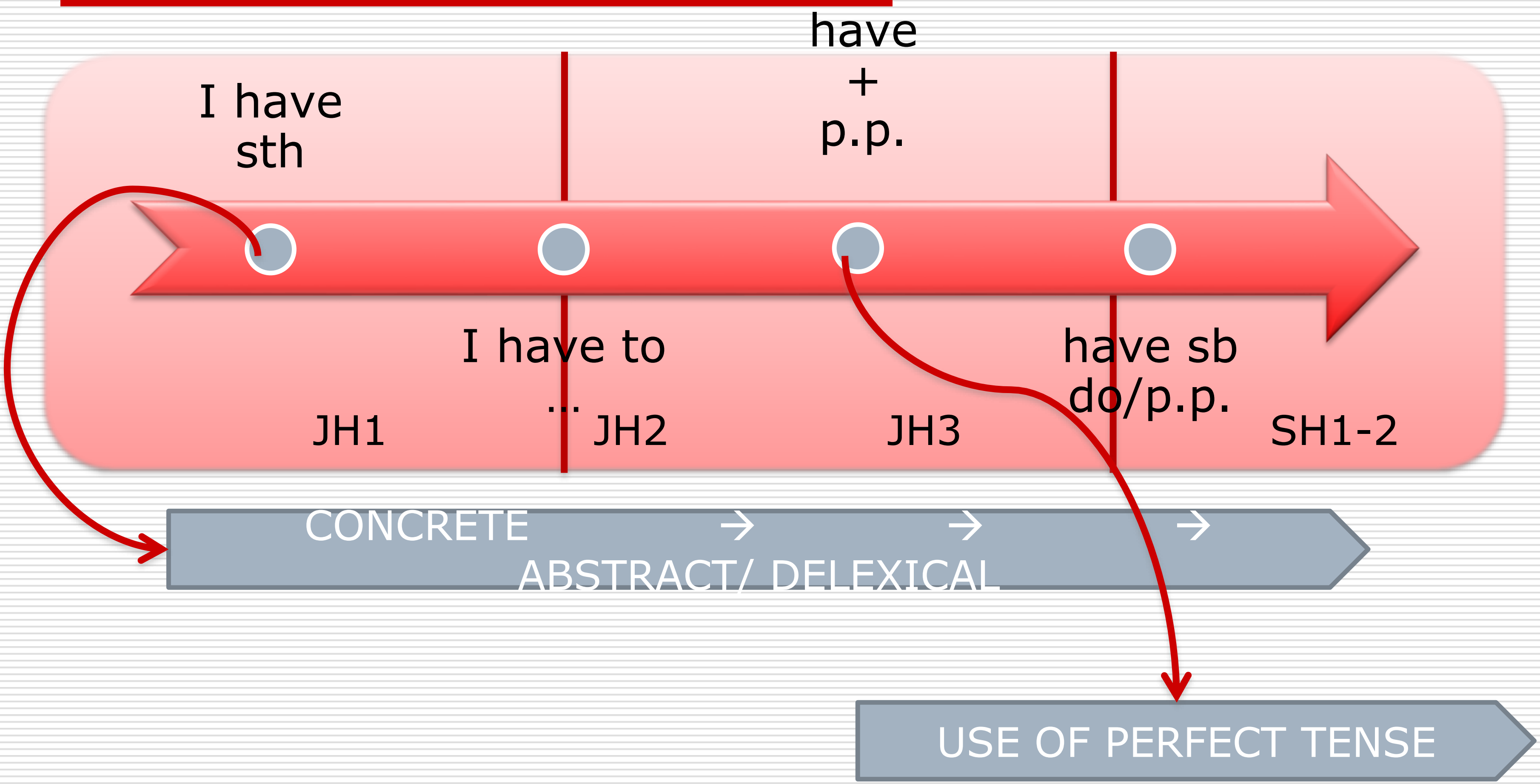
More concrete objects

rice

money

bread

L2 vocabulary profile: Dealing with the gap more efficiently



WORD/POS N-GRAM ANALYSIS

Word *n*-grams (conjunctions)

J1		J2		J3		S1		S2		S3	
Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq
576	#JP# . I	426	**. So I**	644	I do n't	224	I do n't	600	I do n't	235	I do n't
441	. I like	425	I do n't	528	**. So I**	170	#JP# . I	450	I want to	188	I want to
417	**. But I**	408	#JP# . I	527	**. But I**	166	**. But I**	348	**. But I**	166	**. So I**
405	I do n't	378	I will bring	425	#JP# . I	164	**. So I**	343	**. So I**	146	**. But I**
296	**. So I**	347	**. But I**	371	. I was	153	. I was	295	. It is	131	. It was
236	#JP# . #JP#	335	. I like	345	I will bring	147	I want to	291	I ca n't	126	. It is
208	I will bring	275	. I was	336	I want to	139	school festival .	281	. I think	122	in the morning
200	#JP# . But	255	. I will	327	. I like	130	. It was	274	. I was	118	and so on
195	in the morning	233	I want to	283	. I will	114	our school festival	273	**. When I**	116	I ca n't
195	is #JP# .	217	in the morning	283	in the morning	112	. Our class	269	. I like	106	. Our class
189	very much .	193	the morning .	264	**. And I**	109	in the morning	239	a lot of	105	school festival .
178	the morning .	170	very much .	263	I ca n't	103	a lot of	237	and so on	105	so on .
177	. Our class	160	I was very	254	. I do	99	. So ,	235	#JP# . I	103	. Our school
175	Our school festival	159	#JP# . But	228	. I think	98	I was very	229	**. And I**	101	Our school festival
174	. It 's	152	very #JP# .	218	. I usually	90	I will bring	224	. But ,	98	a lot of
173	very #JP# .	151	. So ,	212	the morning .	88	**. And I**	221	**. If I**	95	. I think
171	. I will	149	. It is	209	**. Because I**	88	. I think	216	so on .	94	. I was
161	. Our school	148	. It was	196	do n't have	87	**. Because I**	215	. I do	92	. #JP# is
158	. I do	143	#JP# . So	195	I was very	87	Our school festival	210	. I have	92	I will bring
154	and #JP# .	142	. I want	188	very much .	86	. I like	210	our school festival	92	our school festival
145	. I usually	138	#JP# . #JP#	185	. It is	84	. Our school	208	school festival .	91	the morning .
143	do n't have	138	#JP# . He	179	I usually have	83	I did n't	204	, I will	89	. I do
142	. It is	137	do n't like	171	. I have	82	. I will	204	. So ,	88	do n't have
140	**But I do**	134	**. Because I**	158	#JP# . But	81	. I do	197	in the morning	87	#JP# . I
138	#JP# . Our	134	. I do	158	. It was	79	I ca n't	191	do n't have	86	. So ,
137	. I eat	130	. Our class	158	**. When I**	77	the morning .	182	I will take	85	. I like
134	#JP# . It	128	do n't have	150	. I want	70	**. If I**	181	. It was	85	. I will
131	. #JP# I	127	. I usually	148	. He was	68	So , I	177	. I want	84	**. If I**
129	morning . I	126	. Urashima Taro	147	**. If I**	67	a big earthquake	170	a big earthquake	83	. I want
125	. I 'm	126	I usually have	147	. So ,	66	do n't have	168	**. Because I**	80	**. And I**
124	. I have	123	. Our school	144	#JP# . So	65	, too .	167	**, but I**	78	. I have
120	, too .	122	**. And I**	139	I did n't	65	school festival ,	164	, too .	71	, I will
117	. I #JP#	122	. But ,	136	. One day	61	**, so I**	159	**, so I**	68	. Urashima Taro
115	. Urashima Taro	118	Our school festival	134	a lot of	61	the school festival	152	I think that	65	I could n't
112	. I want	117	. It 's	133	. And he	60	. I have	148	Our school festival	63	. I usually
109	#JP# . So	115	Urashima Taro was	126	, too .	60	. I want	143	**, and I**	63	I usually have
108	is very #JP#	110	a lot of	125	. I 'm	59	#JP# . But	143	. Our school	63	very much .
107	I usually have	106	#JP# . It	125	. So he	58	I could n't	142	. I will	59	. He was
102	#JP# and #JP#	106	I ca n't	123	#JP# . He	58	and #JP# .	137	for breakfast .	59	. It 's
102	every day .	103	is #JP# .	120	and #JP# .	57	, I 'll	136	. For example	58	**. Because I**
101	. I love	102	morning . I	117	. It 's	57	. But ,	132	school festival is	58	. For example
98	**. And I**	100	. I 'm	114	dream . I	55	**, but I**	131	very much .	58	. I 'm
98	rice and #JP#	98	. I think	111	morning . I	55	. It is	127	. It 's	57	**. When I**
98	school festival is	97	it . I	109	and so on	54	#JP# . It	127	I was very	56	**, so I**
97	#JP# . He	97	very happy .	107	do n't like	54	I went to	126	. I 'm	56	I was very
95	#JP# is #JP#	95	I went to	106	, I will	54	very much .	124	So , I	55	for me .
95	do n't like	95	was #JP# .	106	me . I	53	, I was	121	, I was	55	morning . I
88	#JP# in the	93	. I have	104	**, so I**	53	, I will	118	For example ,	51	, #JP# ,
84	I like #JP#	92	#JP# and #JP#	102	breakfast . I	53	. Urashima Taro	115	the morning .	50	. And he
82	bread and milk	92	. I #JP#	101	. Then I	52	very happy .	112	there is a	50	. One day

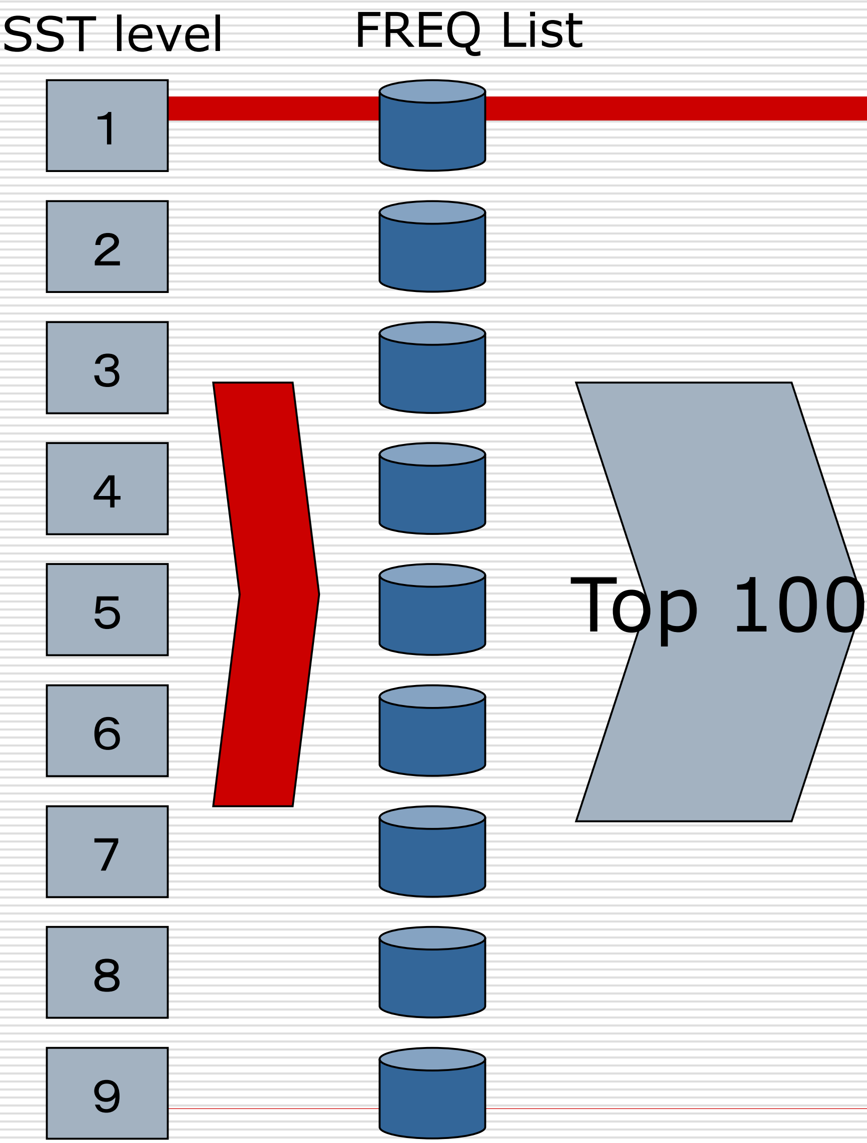
Note: Sentence-initial “but” in blue

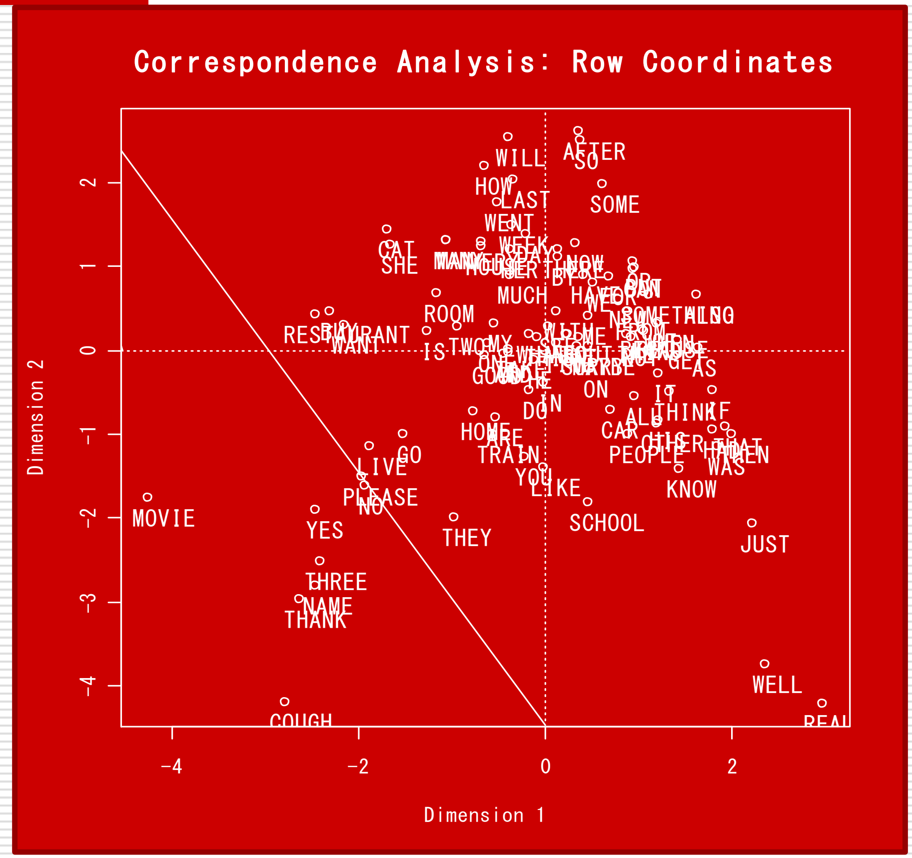
POS *n*-grams (modals =VM)

J1		J2		J3		S1		S2		S3	
Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq	Freq	Seq
1161	. PPIS1 VV0	830	. PPIS1 VV0	1168	NN1 . PPIS1	447	NN1 . PPIS1	1315	NN1 . PPIS1	492	NN1 . PPIS1
692	PPIS1 VV0 NN1	719	NN1 . PPIS1	1033	. PPIS1 VV0	413	**PPIS1 VM VVI**	1259	. PPIS1 VV0	469	**PPIS1 VM VVI**
644	NN1 . PPIS1	679	**PPIS1 VM VVI**	937	**PPIS1 VM VVI**	378	. PPIS1 VV0	1246	**PPIS1 VM VVI**	421	. PPIS1 VV0
640	. APPGE NN1	677	RG JJ .	750	RG JJ .	352	. APPGE NN1	981	II AT NN1	417	II AT NN1
529	#JP# . PPIS1	542	. APPGE NN1	689	JJ NN1 .	341	II AT NN1	934	JJ NN1 .	414	JJ NN1 .
462	RG JJ .	448	JJ NN1 .	645	PPIS1 VD0 XX	333	NN1 NN1 .	806	APPGE NN1 .	375	NN1 NN1 .
424	. CCB PPIS1	432	. RR PPIS1	616	II AT NN1	332	RG JJ .	769	AT1 JJ NN1	335	. APPGE NN1
405	PPIS1 VD0 XX	429	PPIS1 VD0 XX	587	. RR PPIS1	319	APPGE NN1 NN1	737	II APPGE NN1	332	APPGE NN1 NN1
401	**PPIS1 VM VVI**	410	. PPIS1 VVD	583	VD0 XX VVI	302	VBDZ RG JJ	722	. CS PPIS1	329	II APPGE NN1
322	PPIS1 VV0 #JP#	399	PPIS1 VV0 NN1	581	APPGE NN1 .	281	. PPIS1 VVD	698	AT NN1 .	302	APPGE NN1 .
321	APPGE NN1 NN1	387	VBDZ RG JJ	558	VBDZ RG JJ	264	JJ NN1 .	667	NN1 , PPIS1	299	**VM XX VVI**
316	. PPH1 VBZ	382	#JP# . PPIS1	555	. CS PPIS1	257	APPGE NN1 .	661	**, PPIS1 VM**	271	AT NN1 .
314	VBZ RG JJ	381	**. PPIS1 VM**	549	. CCB PPIS1	248	AT NN1 .	660	**VM XX VVI**	262	RG JJ .
299	NN1 NN1 .	365	NN1 NN1 .	536	AT NN1 .	246	II APPGE NN1	654	APPGE NN1 NN1	251	**PPIS1 VM XX**
288	NN1 CC NN1	352	. CCB PPIS1	508	. APPGE NN1	226	PPIS1 VD0 XX	646	. APPGE NN1	251	VBDZ RG JJ
287	. RR PPIS1	349	VD0 XX VVI	499	. PPIS1 VVD	224	. CS PPIS1	635	NN1 NN1 .	245	NN1 CC NN1
284	NN1 #JP# .	321	APPGE NN1 .	477	**. PPIS1 VM**	206	AT1 JJ NN1	610	. PPIS1 VVD	245	VV0 TO VVI
284	VD0 XX VVI	307	VBZ RG JJ	461	**VM XX VVI**	198	VD0 XX VVI	609	PPIS1 VD0 XX	238	**, PPIS1 VM**
273	VV0 NN1 .	292	AT NN1 .	458	**PPIS1 VM XX**	193	. RR PPIS1	606	VV0 TO VVI	237	PPIS1 VD0 XX
259	JJ #JP# .	282	APPGE NN1 NN1	457	PPIS1 VV0 NN1	192	AT1 NN1 IO	593	NN1 . CS	231	. CS PPIS1
252	NN1 . CCB	275	. PPIS1 VBDZ	436	NN1 . CCB	189	**, PPIS1 VM**	579	, PPIS1 VV0	230	. PPHS1 VVD
236	#JP# . #JP#	275	NN1 . CCB	422	II APPGE NN1	182	PPIS1 VV0 TO	577	PPIS1 VV0 NN1	230	PPIS1 VV0 TO
234	**. PPIS1 VM**	275	PPIS1 VV0 TO	416	#JP# . PPIS1	182	**VM XX VVI**	568	PPIS1 VV0 TO	221	AT1 JJ NN1
232	#JP# . APPGE	274	II AT NN1	409	PPIS1 VV0 TO	178	NN1 , PPIS1	562	RG JJ .	217	NN1 . RR
229	APPGE NN1 .	269	JJ . PPIS1	393	. PPHS1 VVD	176	VV0 TO VVI	550	VD0 XX VVI	214	. RR ,
221	JJ NN1 .	267	. PPH1 VBZ	385	NN1 . CS	172	**PPIS1 VM XX**	541	**PPIS1 VM XX**	213	**. PPIS1 VM**
221	NNT1 . PPIS1	246	NN1 CC NN1	380	NN1 . RR	171	. CCB PPIS1	513	**. PPIS1 VM**	205	NN1 , PPIS1
212	VBZ #JP# .	244	JJ #JP# .	380	NN1 NN1 .	169	**. PPIS1 VM**	502	NN1 CC NN1	202	. RR PPIS1
207	#JP# . CCB	238	. PPHS1 VVD	376	VV0 TO VVI	169	NN1 CC NN1	482	VBZ RG JJ	199	AT1 NN1 IO
203	II AT NNT1	237	II AT NNT1	371	. PPIS1 VBDZ	168	PPIS1 VV0 NN1	469	AT1 NN1 IO	197	VD0 XX VVI
197	NN1 CC #JP#	235	NN1 . RR	368	. PPIS1 RR	167	NN1 . CCB	442	. RR PPIS1	195	NN1 . CS
194	. PPIS1 VBM	228	**VM XX VVI**	364	VBZ RG JJ	165	AT NN1 NN1	441	NN1 . RR	185	. PPH1 VBZ
192	NN1 NN1 VBZ	228	VV0 TO VVI	342	CS PPIS1 VV0	165	NN1 . CS	437	CS PPIS1 VV0	179	VBZ RG JJ
191	RG DA1 .	223	II APPGE NN1	341	JJ . PPIS1	164	#JP# . PPIS1	434	NN1 . CCB	178	CS PPIS1 VV0
187	AT NNT1 .	210	AT NNT1 .	334	AT1 JJ NN1	161	, PPIS1 VV0	424	. PPH1 VBZ	177	. PPIS1 VVD
186	NN1 . APPGE	207	**PPIS1 VM XX**	331	NN1 CC NN1	161	VBZ RG JJ	418	VBDZ RG JJ	170	PPIS1 VV0 NN1
185	. PPIS1 RR	193	. CS PPIS1	316	II AT NNT1	154	VVD TO VVI	407	. RR ,	169	NN1 . PPH1
182	AT1 NNT1 .	190	NNT1 . PPIS1	314	AT1 NN1 .	153	. PPIS1 VBDZ	405	JJ NN1 ,	164	NN1 , NN1
179	JJ . PPIS1	189	JJ . CCB	312	. PPH1 VBZ	150	NN1 . RR	390	NN1 IO NN1	162	, PPIS1 VV0
166	CC #JP# .	184	NN1 . APPGE	298	NN1 , PPIS1	149	. RR ,	386	AT NN1 ,	161	NN1 II AT
164	NN1 VBZ #JP#	184	NN1 CC #JP#	289	II NN1 .	146	II AT NNT1	361	. PPHS1 VVD	159	AT1 NN1 .
163	APPGE NN1 VBZ	179	II NN1 .	288	APPGE NN1 NN1	146	JJ . PPIS1	359	. CCB PPIS1	159	NN1 . CCB
161	#JP# PPIS1 VV0	176	. PPIS2 VVD	279	NN1 II AT	145	NN1 NN1 ,	357	JJ . PPIS1	156	II AT NNT1
160	APPGE NN1 #JP#	175	. PPIS1 RR	276	**, PPIS1 VM**	143	AT1 NN1 .	354	II NN1 .	155	. PPIS1 RR
160	AT NN1 .	175	NN1 II AT	269	. CC PPIS1	140	, PPIS1 VVD	343	AT1 NN1 .	154	II NN1 .
160	VV0 NN1 CC	174	RG DA1 .	264	PPIS1 RR VV0	132	CS PPIS1 VV0	341	NN1 . PPH1	153	AT NN1 ,
158	. PPIS1 VD0	172	#JP# . CCB	257	NN1 . PPHS1	130	. PPH1 VBDZ	330	TO VVI NN1	150	NN1 IO NN1
155	RR PPIS1 VV0	164	. PPIS1 VBM	255	NN1 . CC	128	APPGE NN1 VVD	326	, PPIS1 VVD	149	NN1 . PPHS1
154	CCB PPIS1 VD0	164	. RR ,	255	PPIS1 RR VH0	126	TO VVI NN1	325	APPGE NN1 ,	149	NN1 NN1 ,
152	NN1 VBZ RG	163	NN1 #JP# .	254	. PPIS1 VD0	125	. PPHS1 VVD	321	, CS PPIS1	148	. CCB PPIS1
148	#JP# NN1 #JP#	160	AT1 JJ NN1	244	, PPIS1 VV0	122	NN1 II AT	320	NN1 II AT	146	VVD TO VVI
145	VD0 XX VHI	159	NN1 . CS	235	AT NNT1 .	119	VDD XX VVI	314	**VM VVI RP**	133	JJ . PPIS1

MULTIVARIATE ANALYSIS (CLUSTERING)

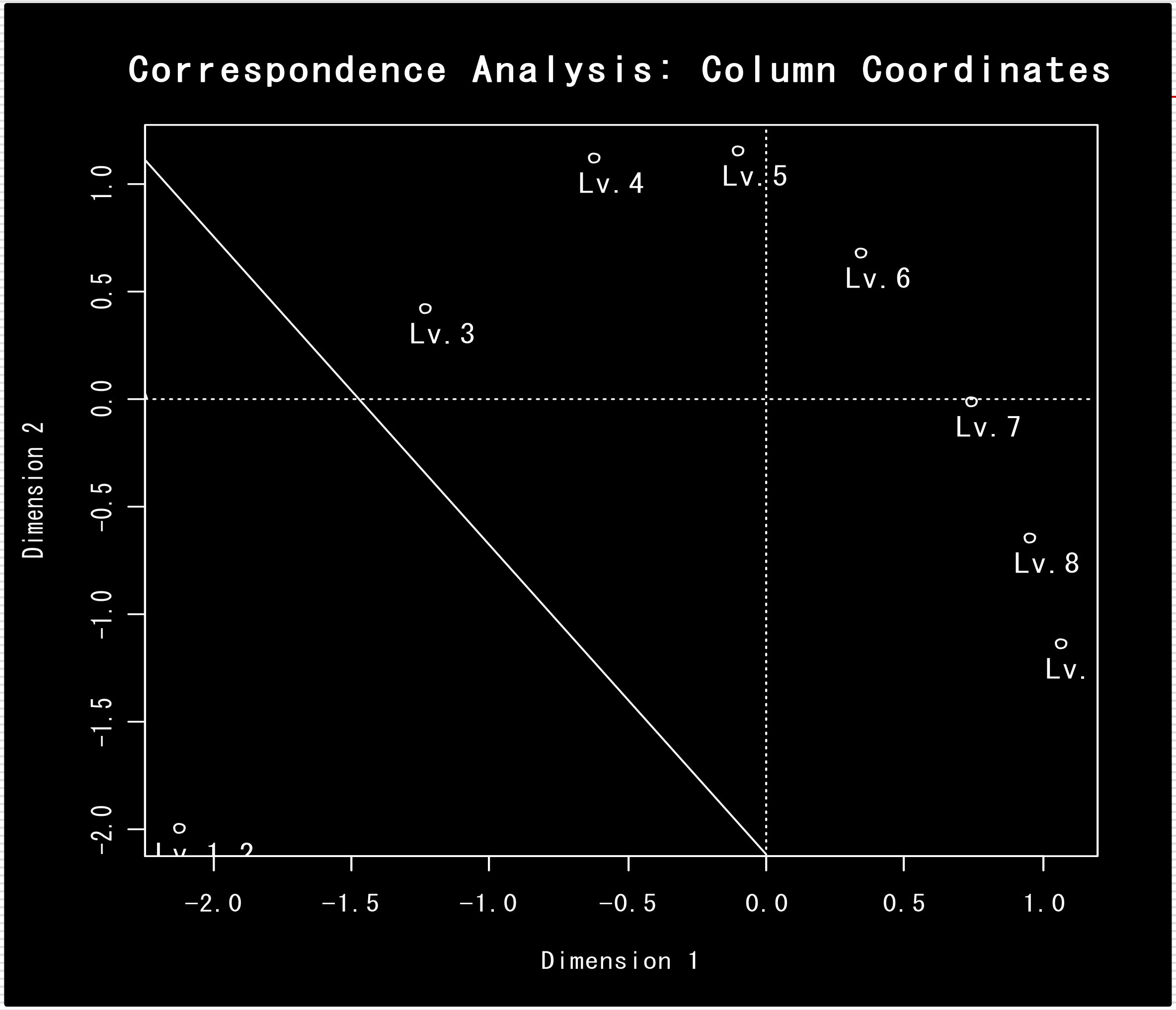
Analysis of Spoken Corpus (NICT JLE)





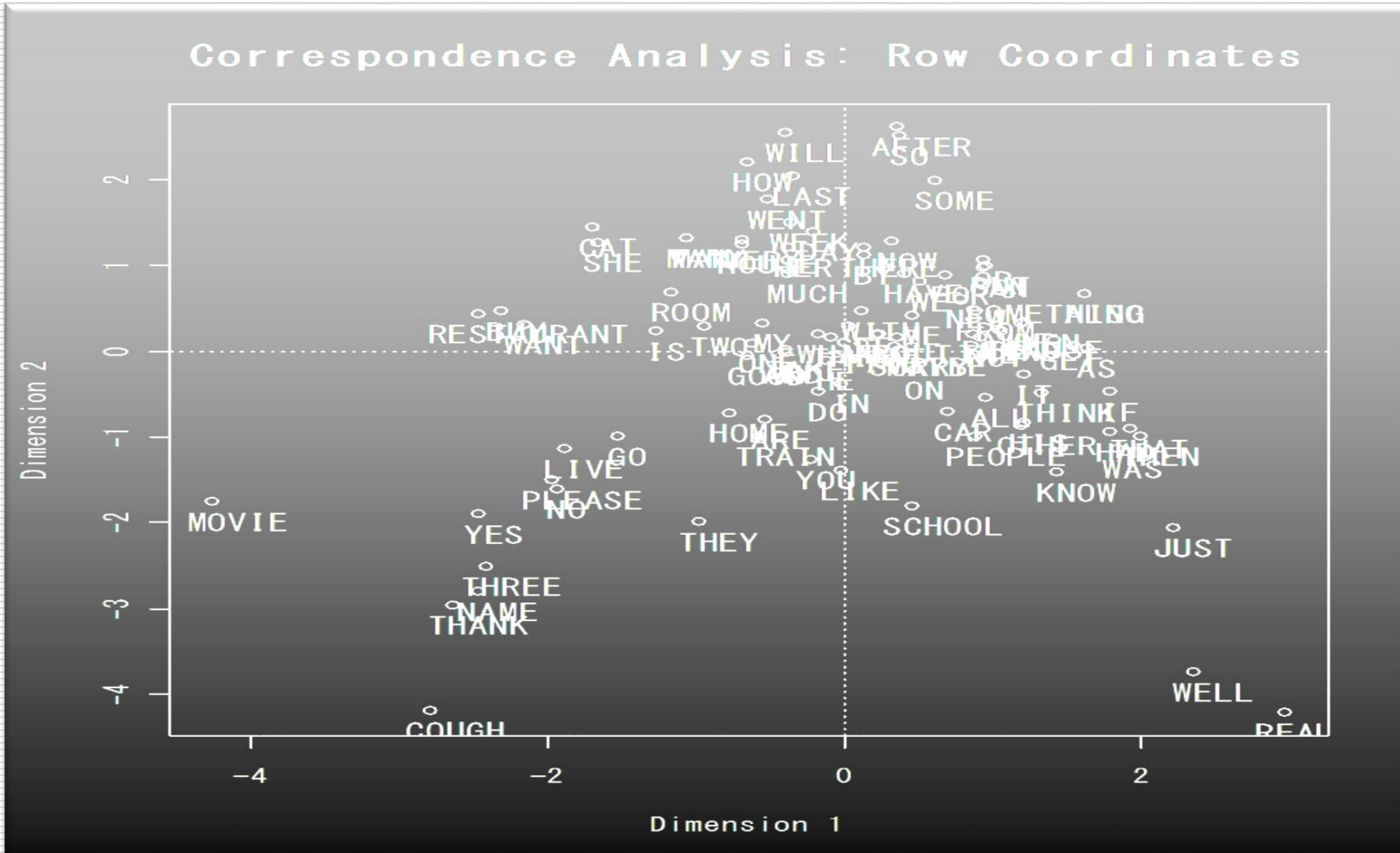
Data summarization

Correspondence Analysis

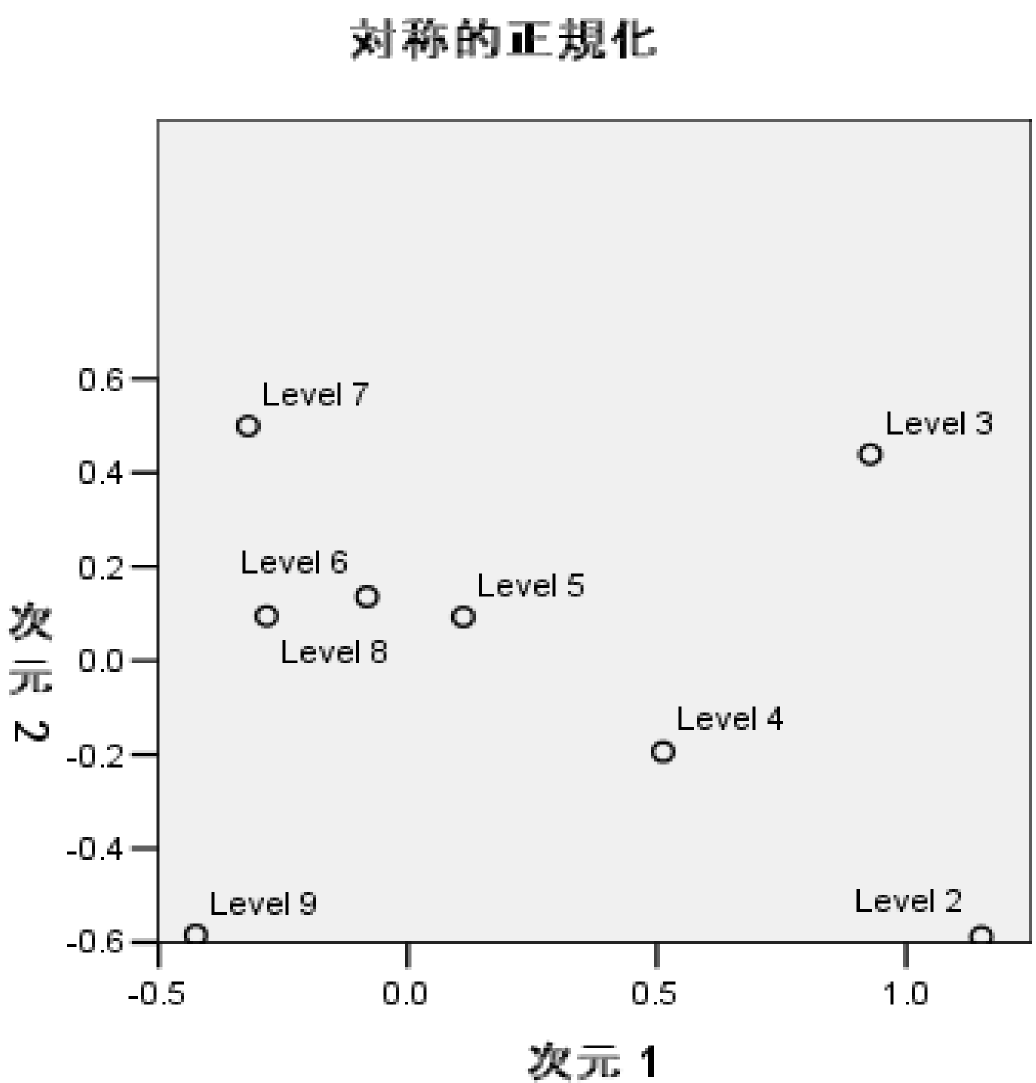


- The most frequent 100 wo
can serve as a useful crite
feature for distinguishing o
level from another.
-

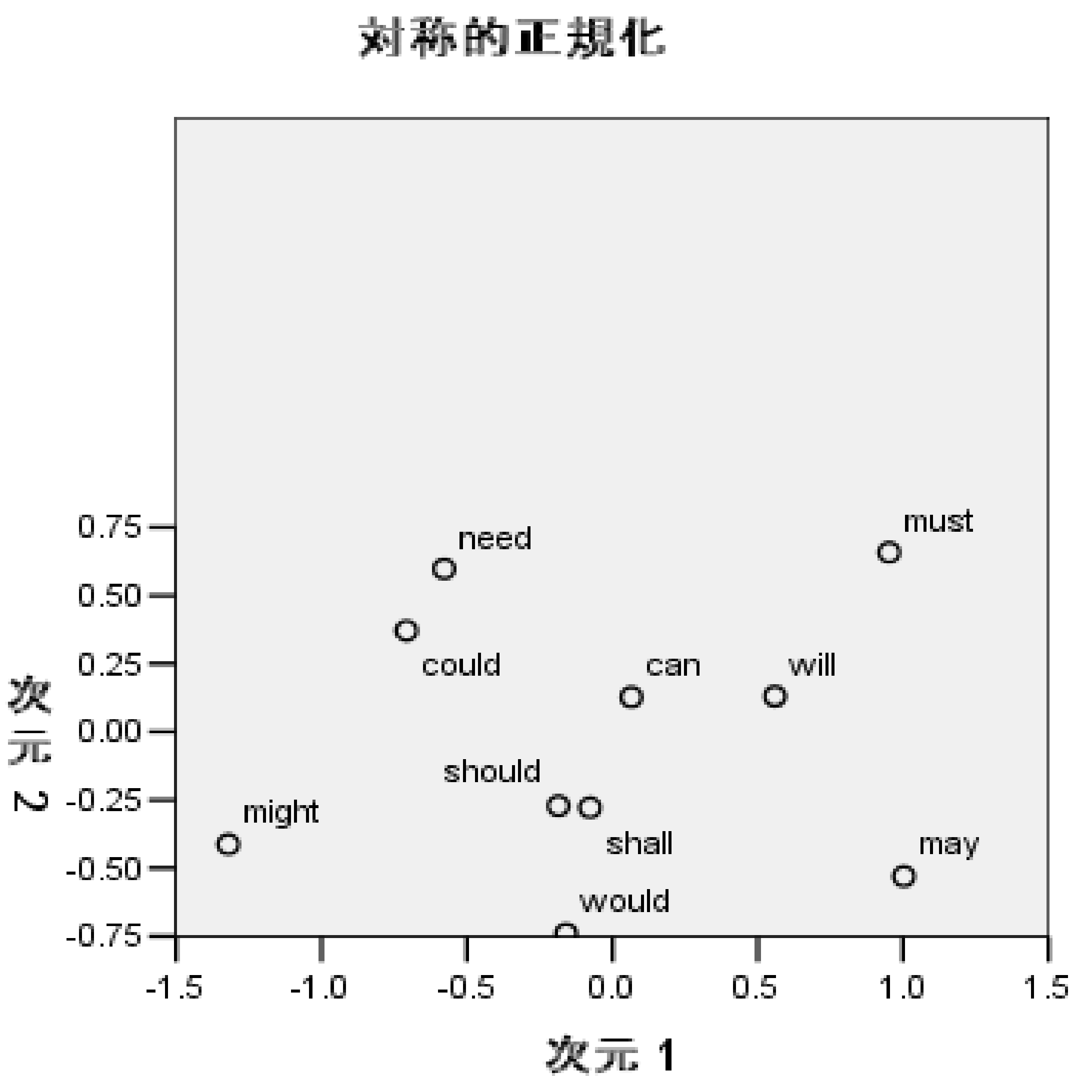
Correspondence Analysis



The use of modal auxiliaries across different proficiency

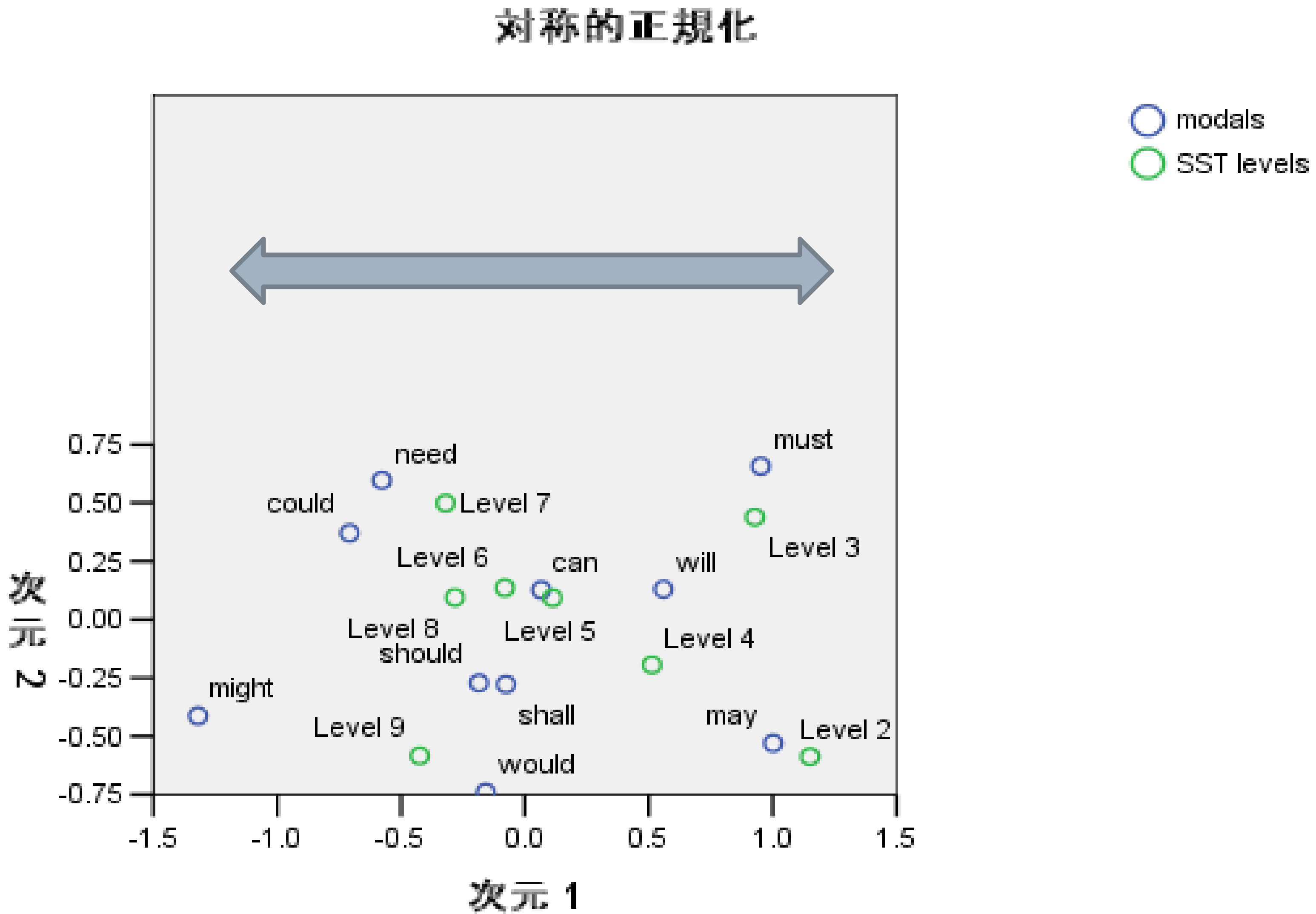


The use of modal auxiliaries across different proficiency



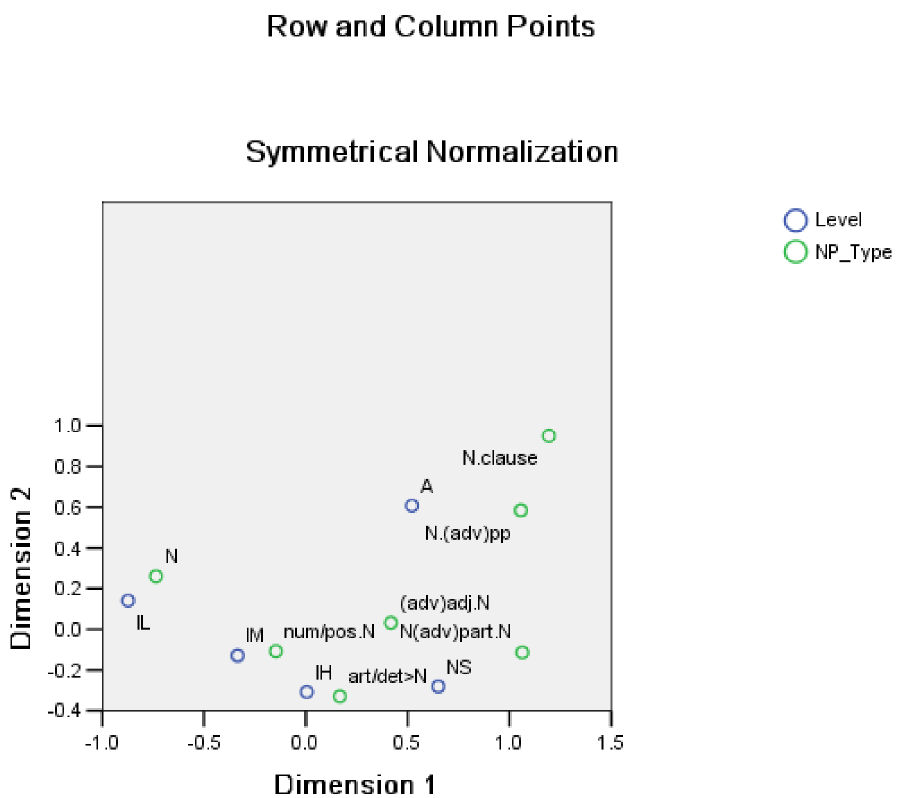
行ポイントと列ポイント

The use of modal auxiliaries across different proficiency



Kaneko (2006): NP structures

NICT-JLE

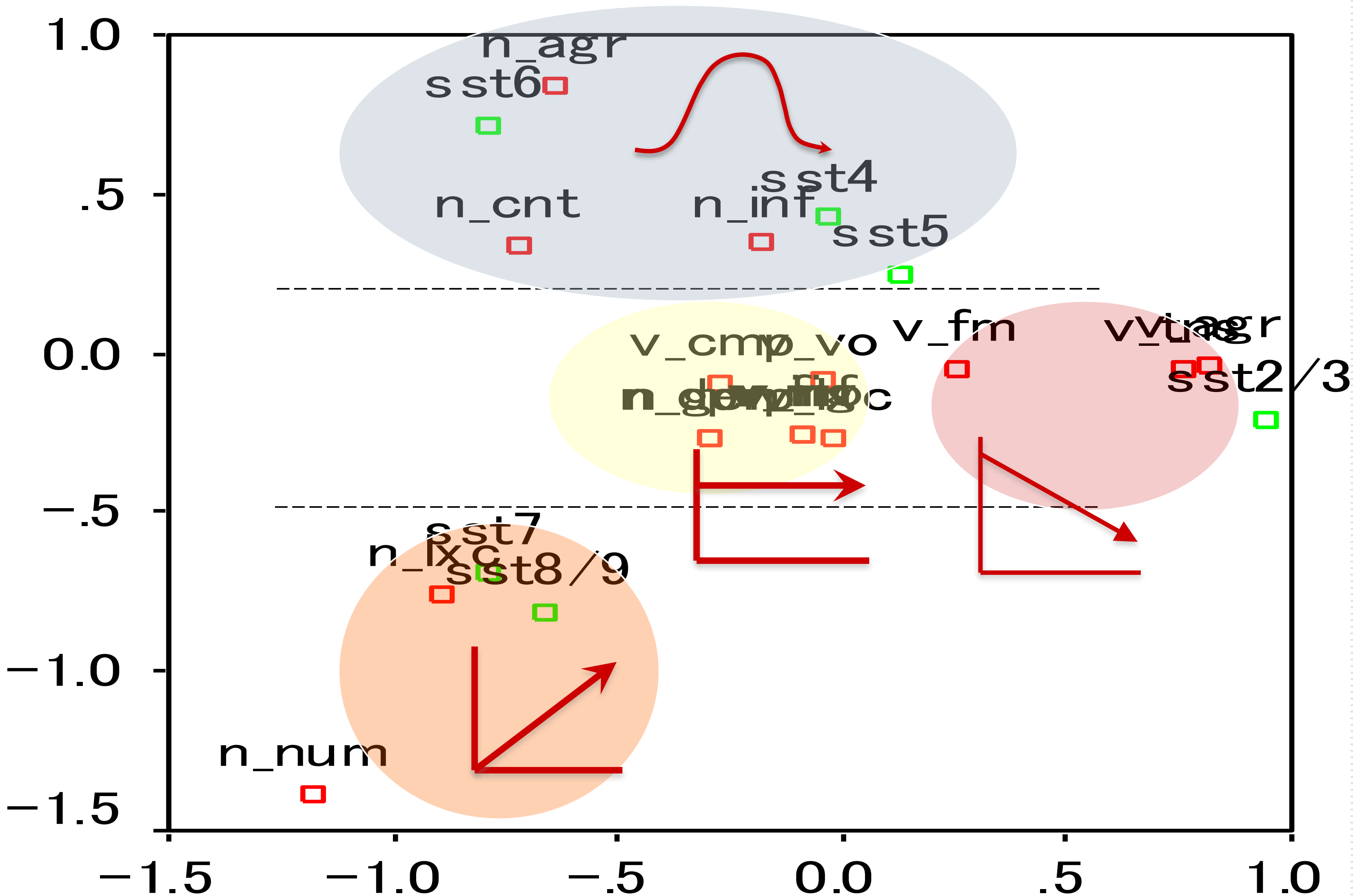


Proficiency levels

NP types:
- N
- num/possessive + N
- det + N
- N (adv)part + N
- (adv) adj + N
- N + (adv) + PP
- N + clause

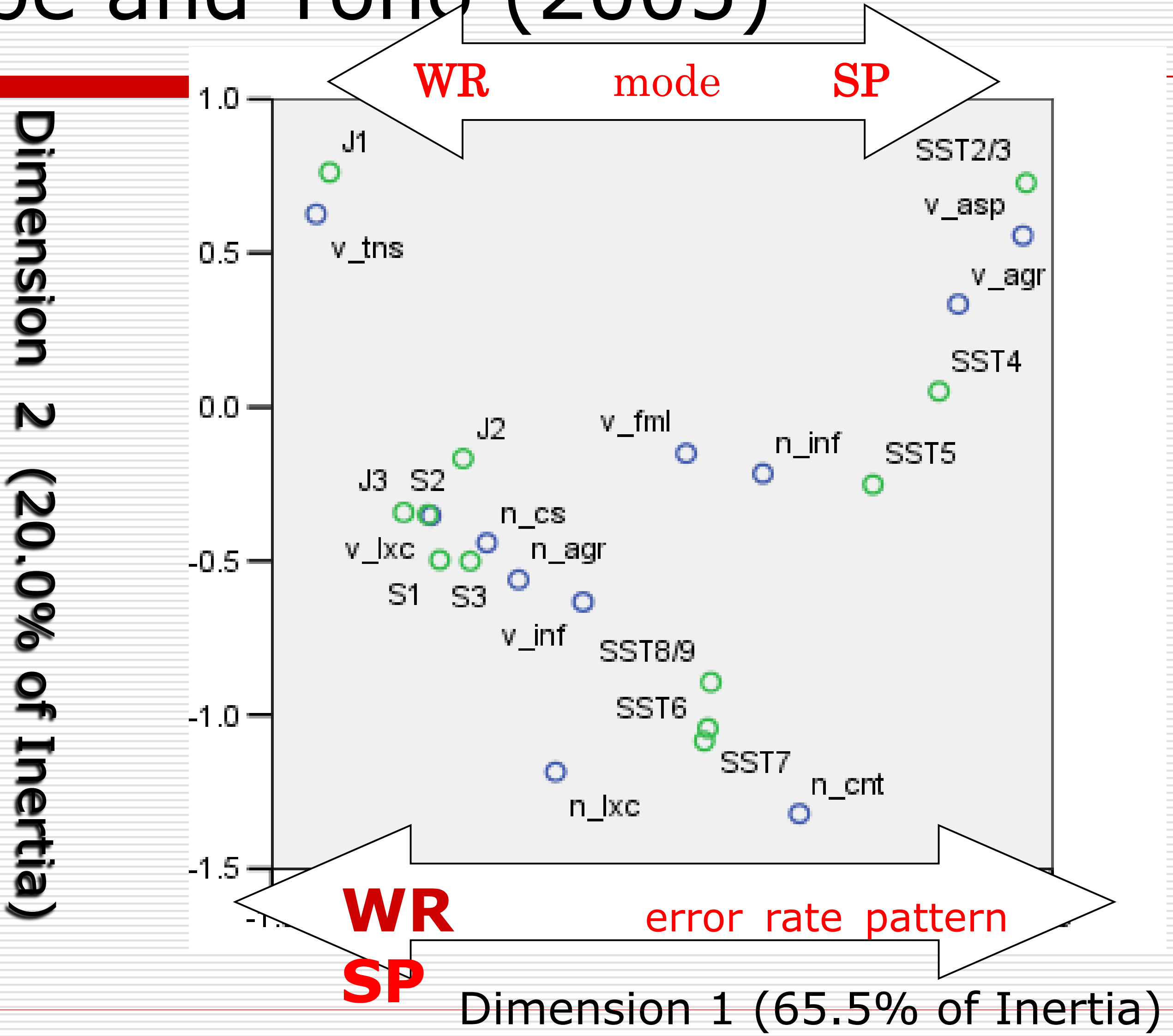
Learner Levels:
- IL = Intermediate (low
- IM = Intermediate (mid
- IH = Intermediate (hig
- A = advanced
- NS = native speaker

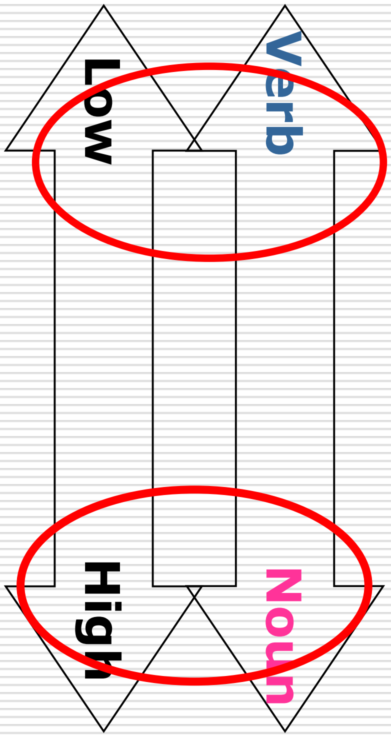
Error freq's & distributions



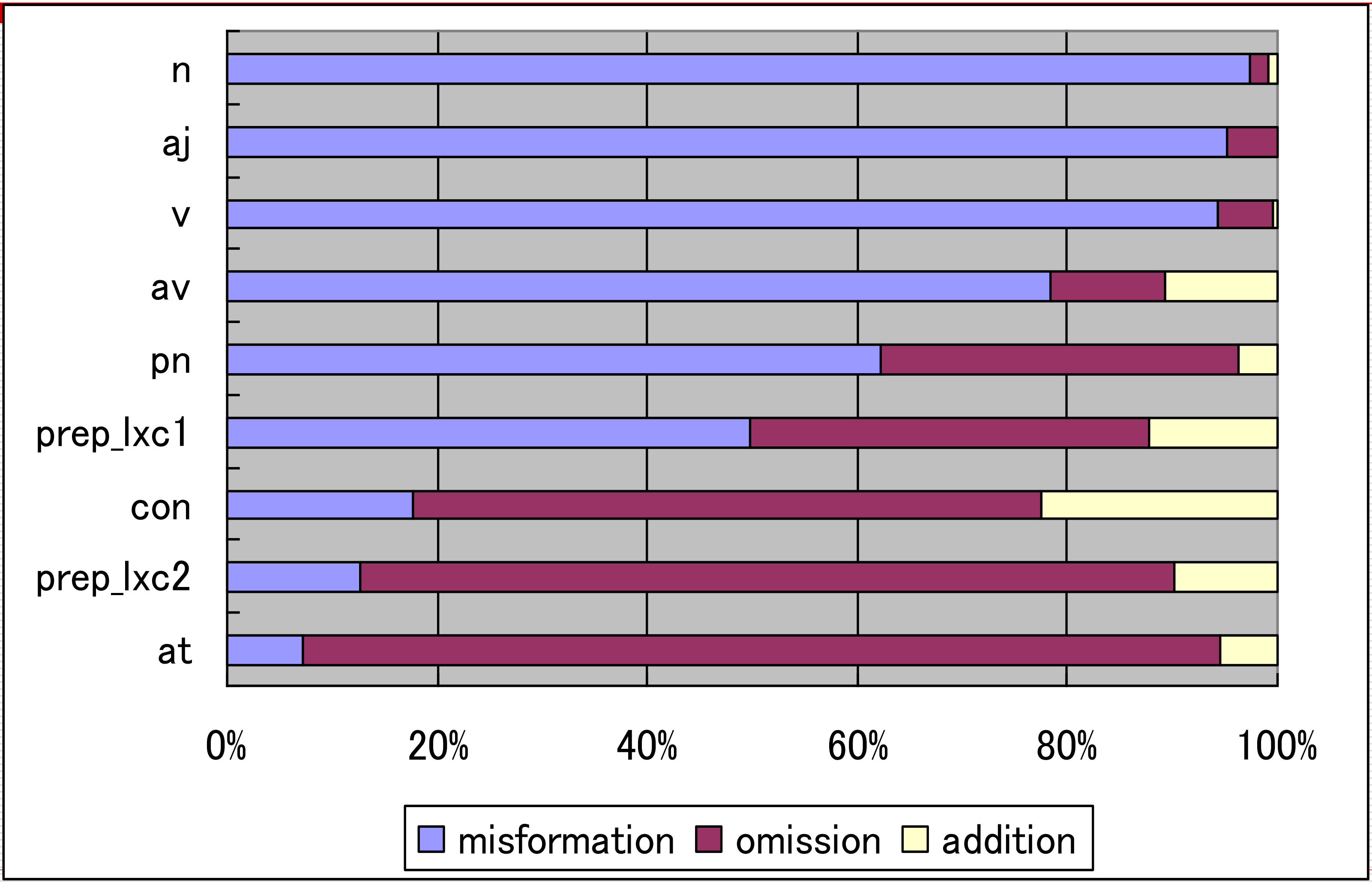
Dimension 1（64．2%） Dimension 2（寄与率20．6%）

Abe and Tono (2005)





Error types across POS (Abe & Tono 2005)

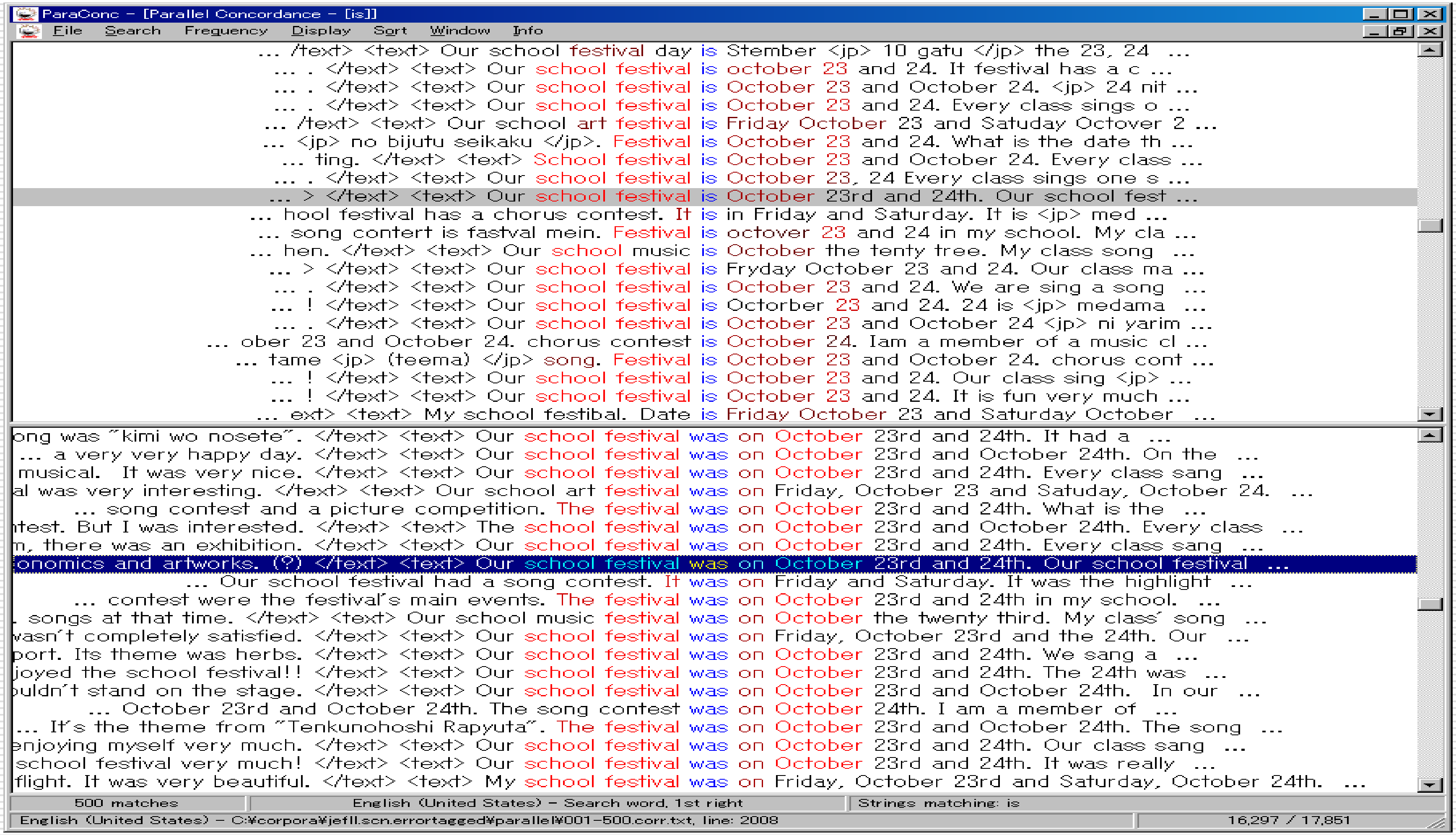


AUTOMATIC ERROR IDENTIFICATION

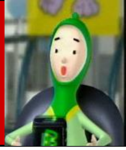
Automatic identification of learner errors

- 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

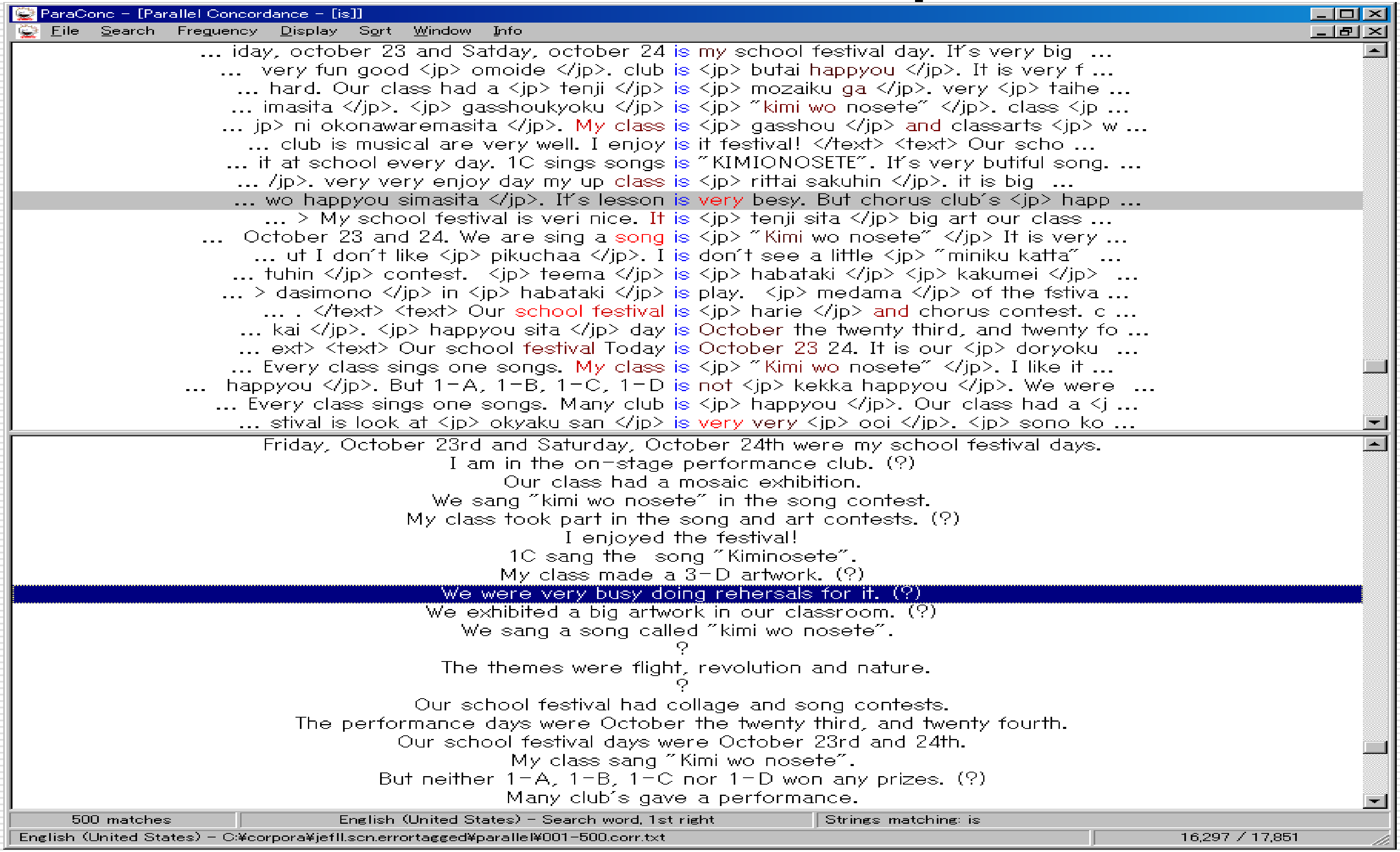
Parallel concordancing



Upper pain: original vs. lower pain: corrected



Errors involved in copula "*be*"



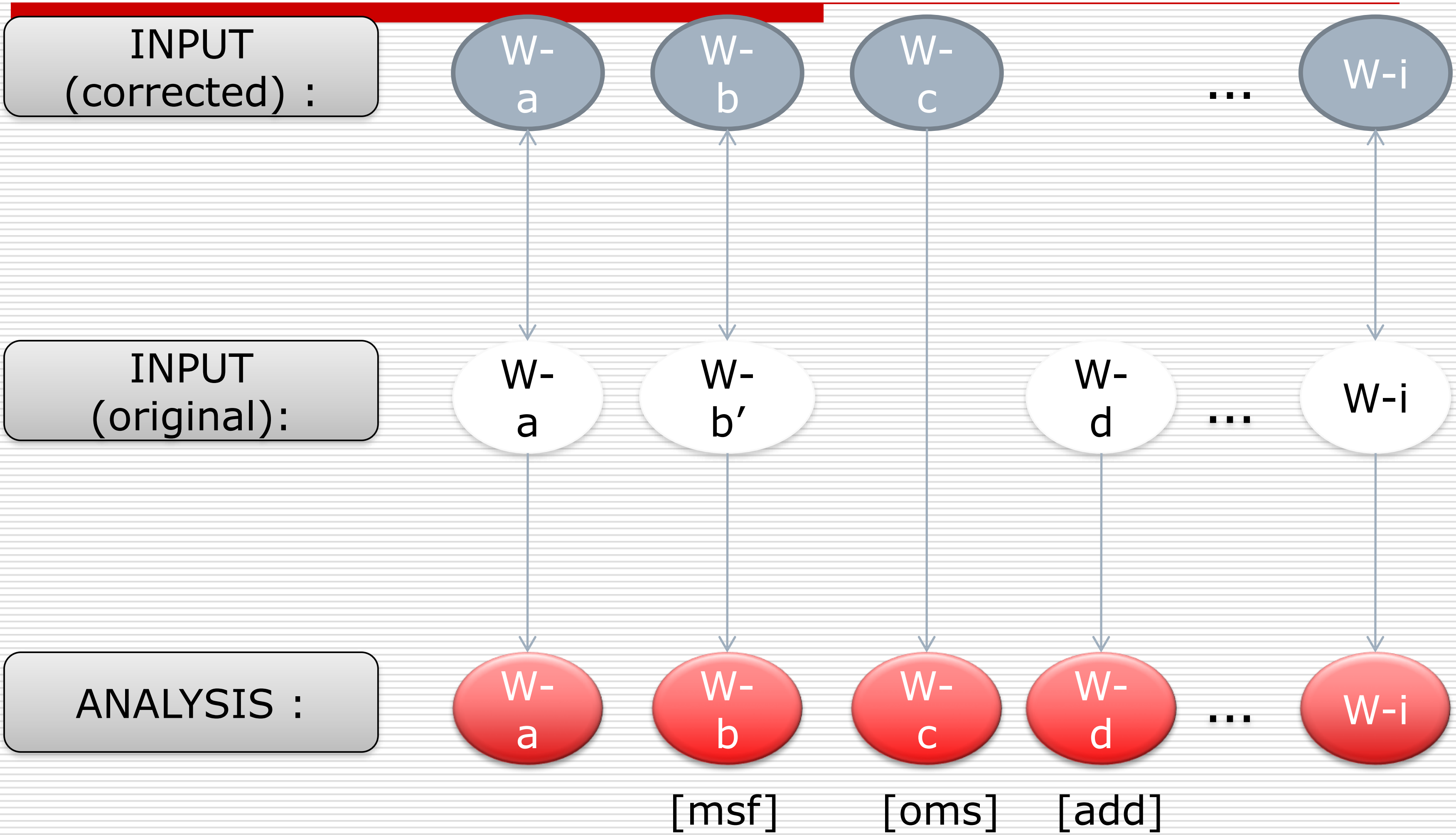
ParaConc – [Parallel Concordance – [is]]
File Search Frequency Display Sort Window Info

... 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

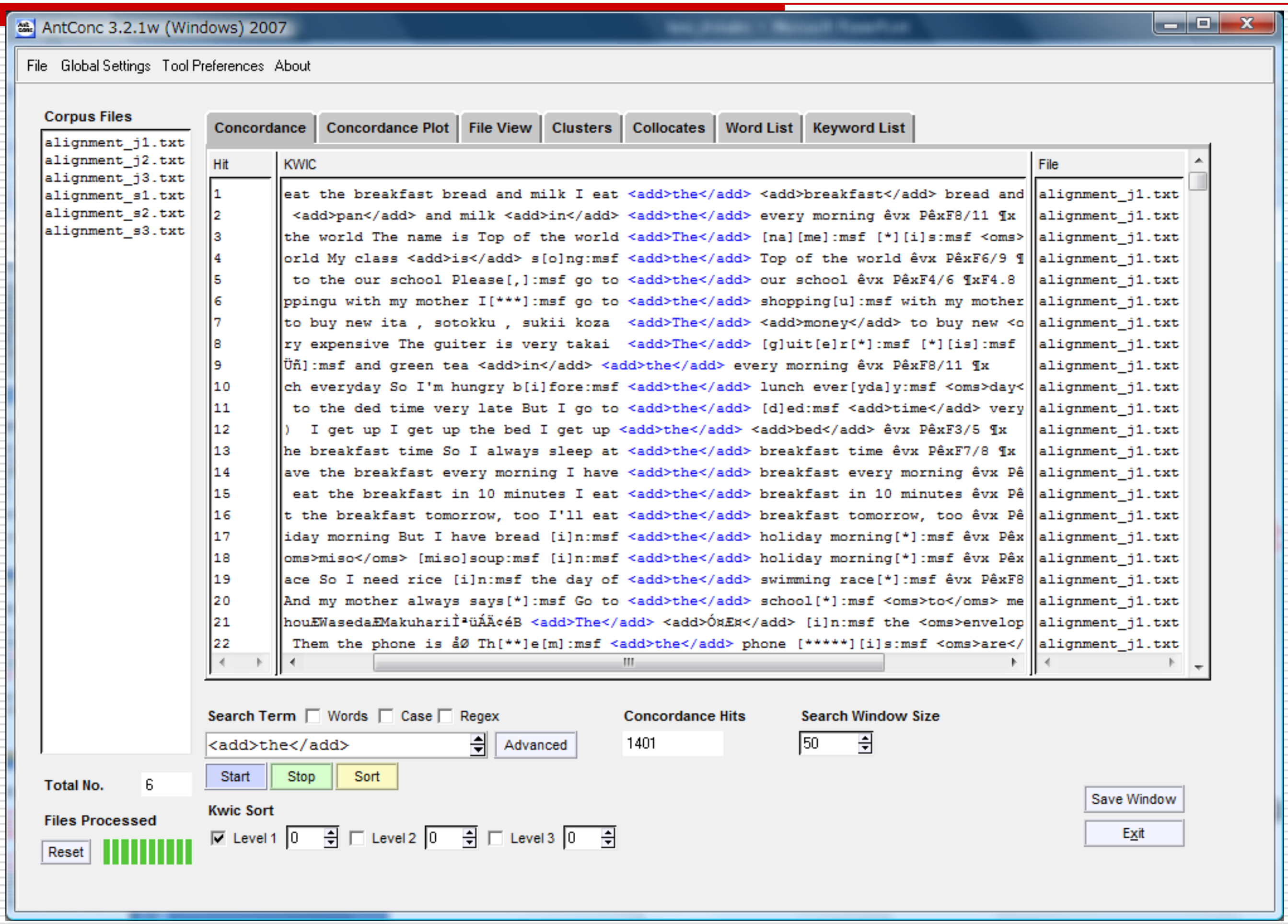
O: mother is cook very well ← original sentence

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

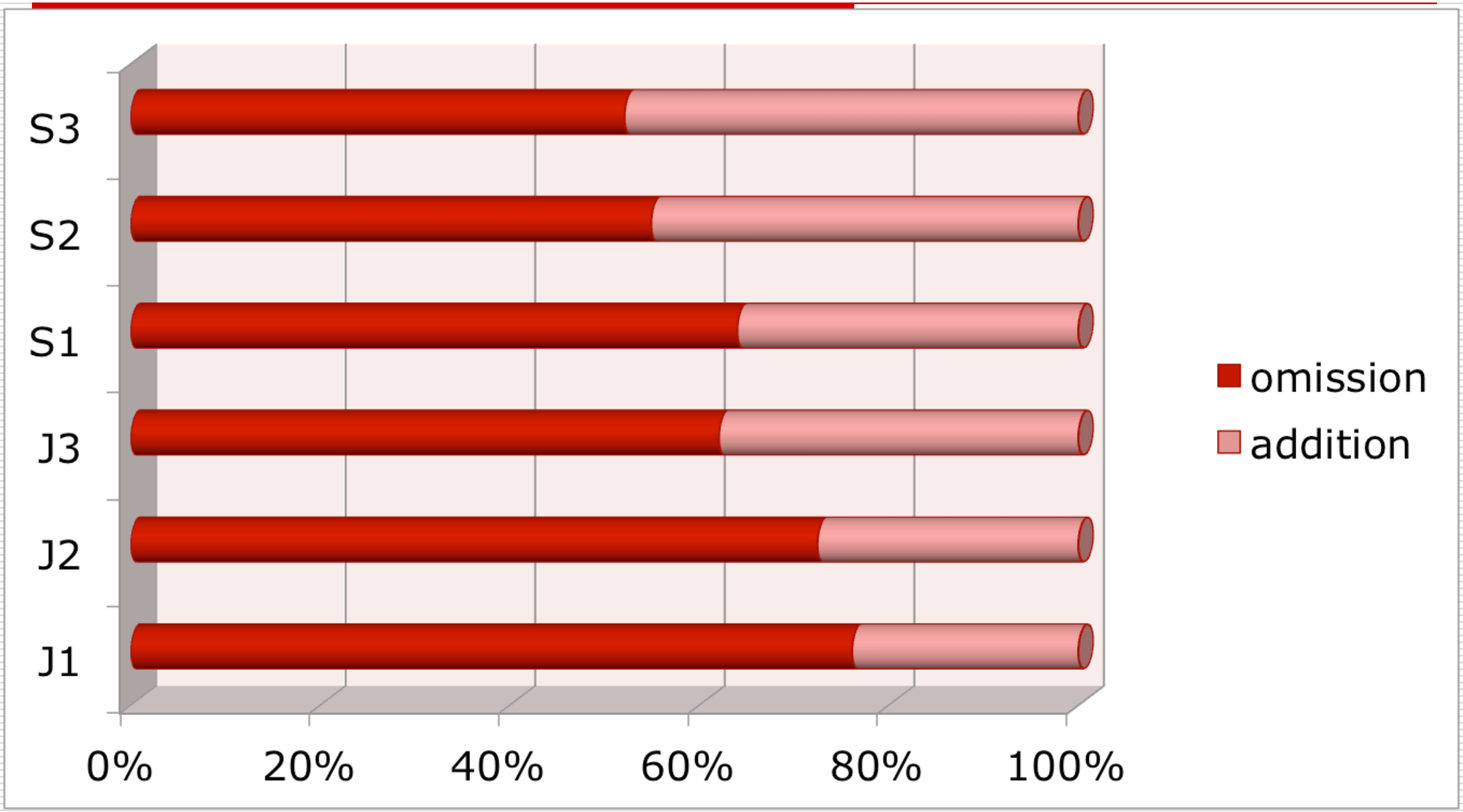
- Correspondence ratio:
  - Word level : 3/5
  - Character level : 3.80/5(76%)

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

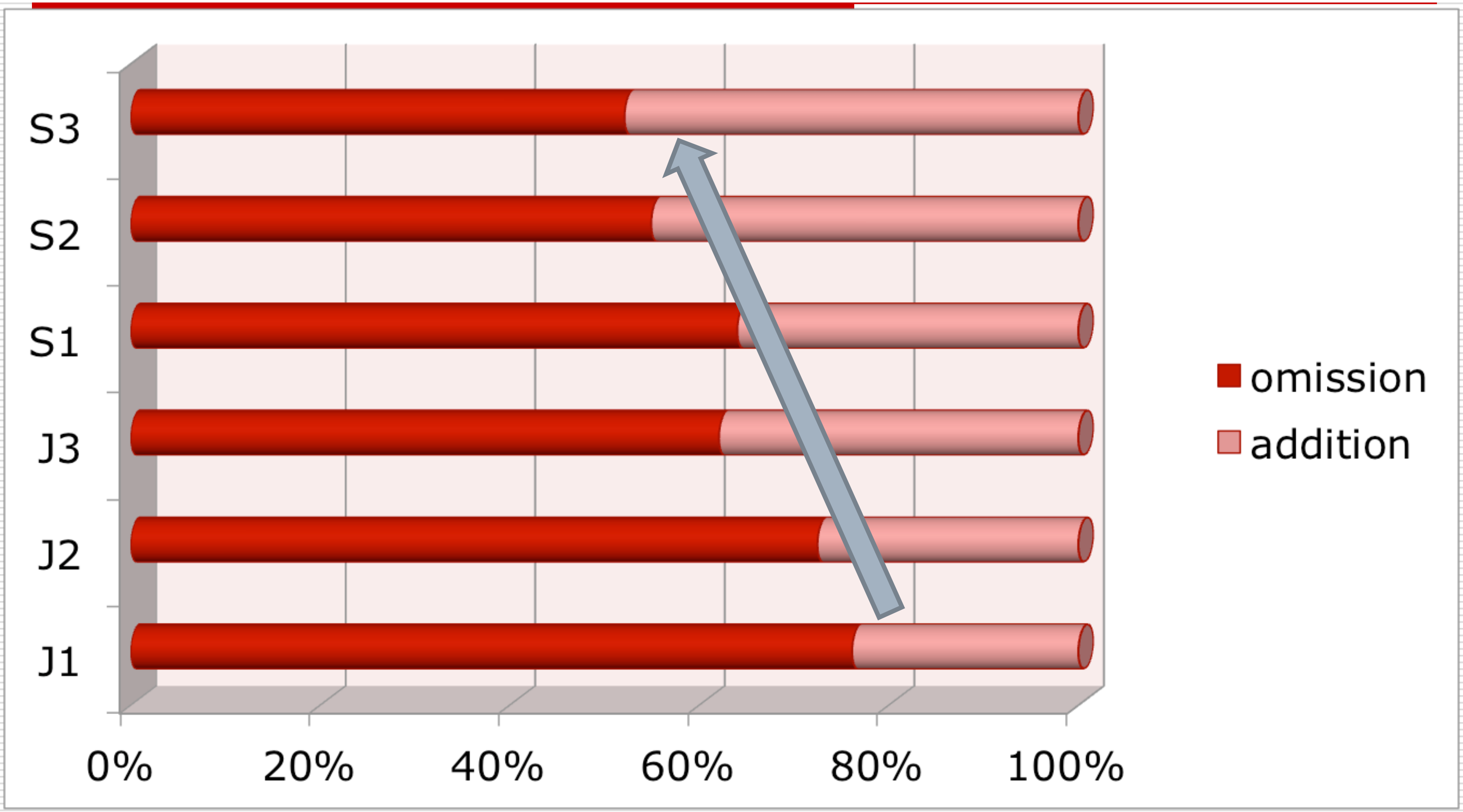
Looking for criterial features



Distributions of error types

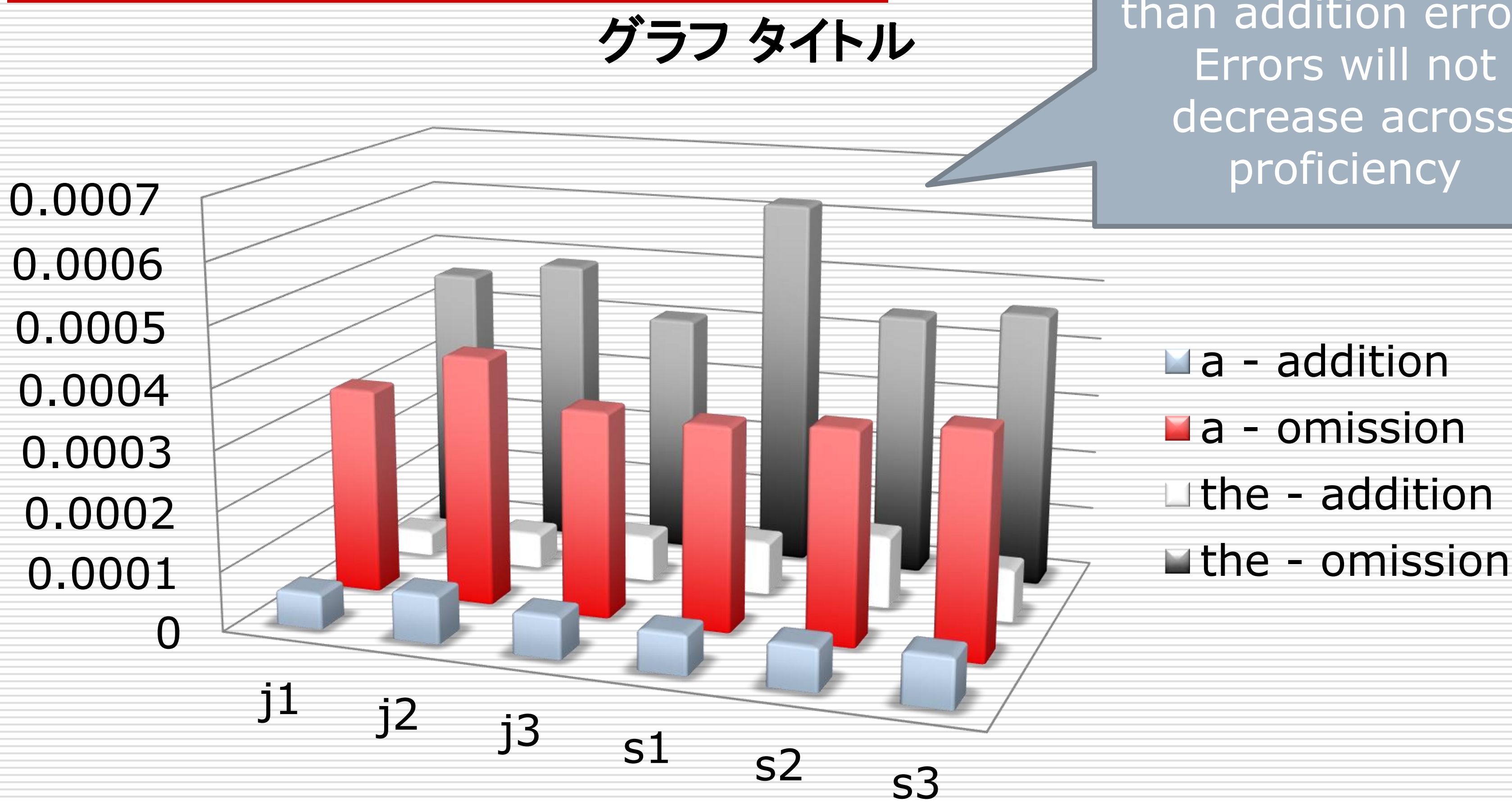


Distributions of error types



Article errors

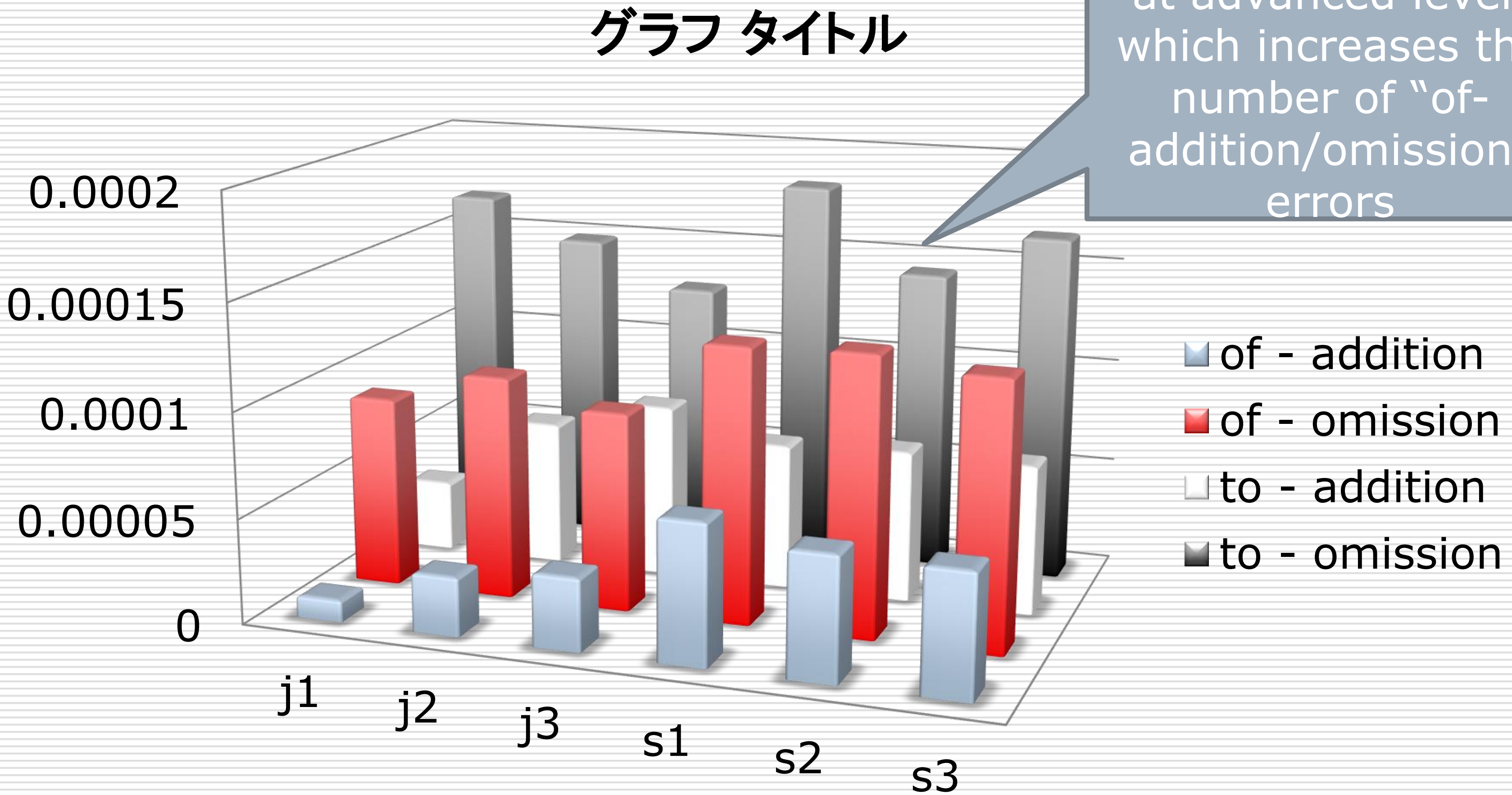
Omission errors are far more common than addition errors. Errors will not decrease across proficiency



Omission errors are significantly more frequent than addition errors.

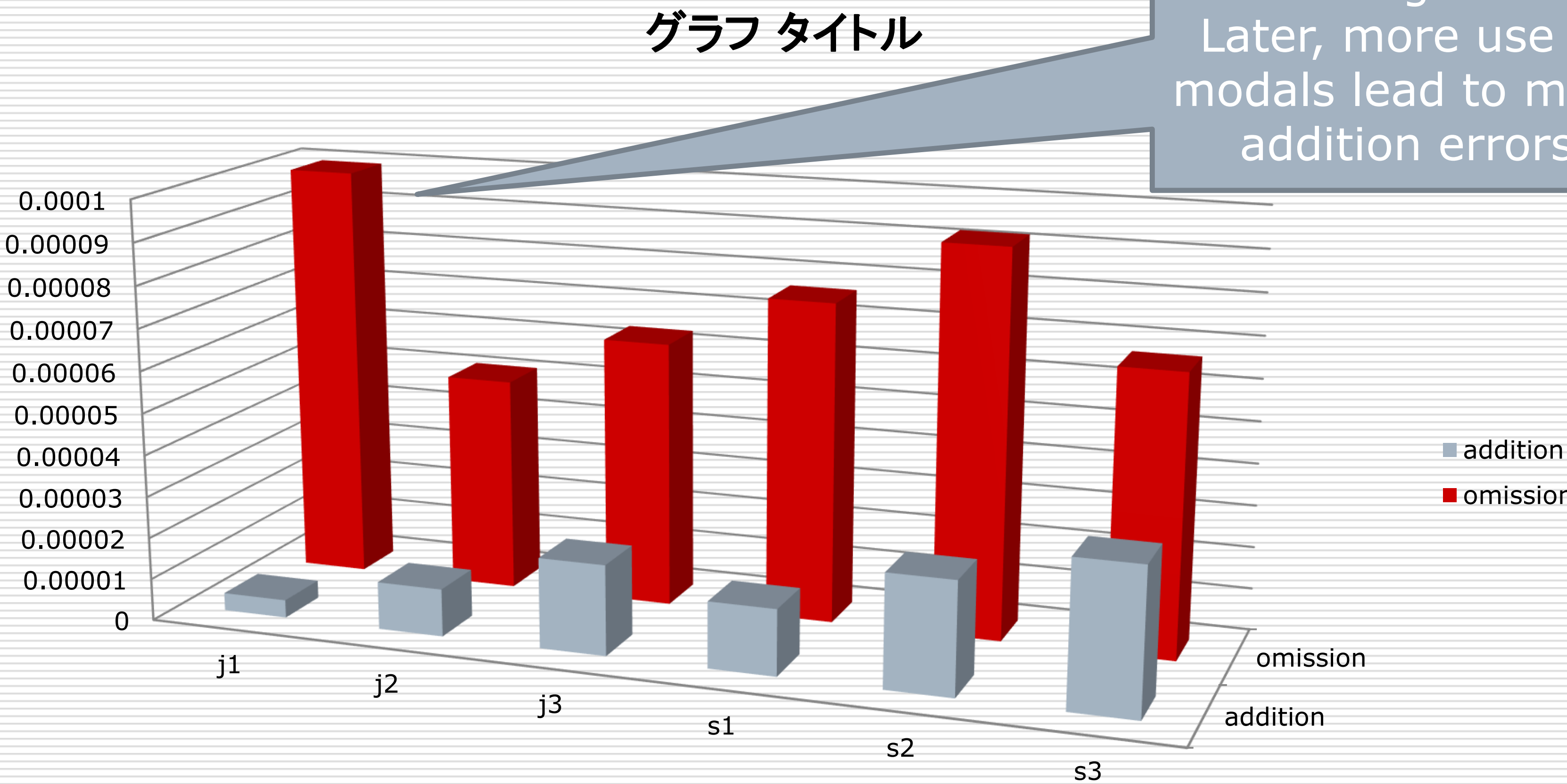
Preposition errors (to/of)

More nominalization at advanced level, which increases the number of "of-addition/omission" errors

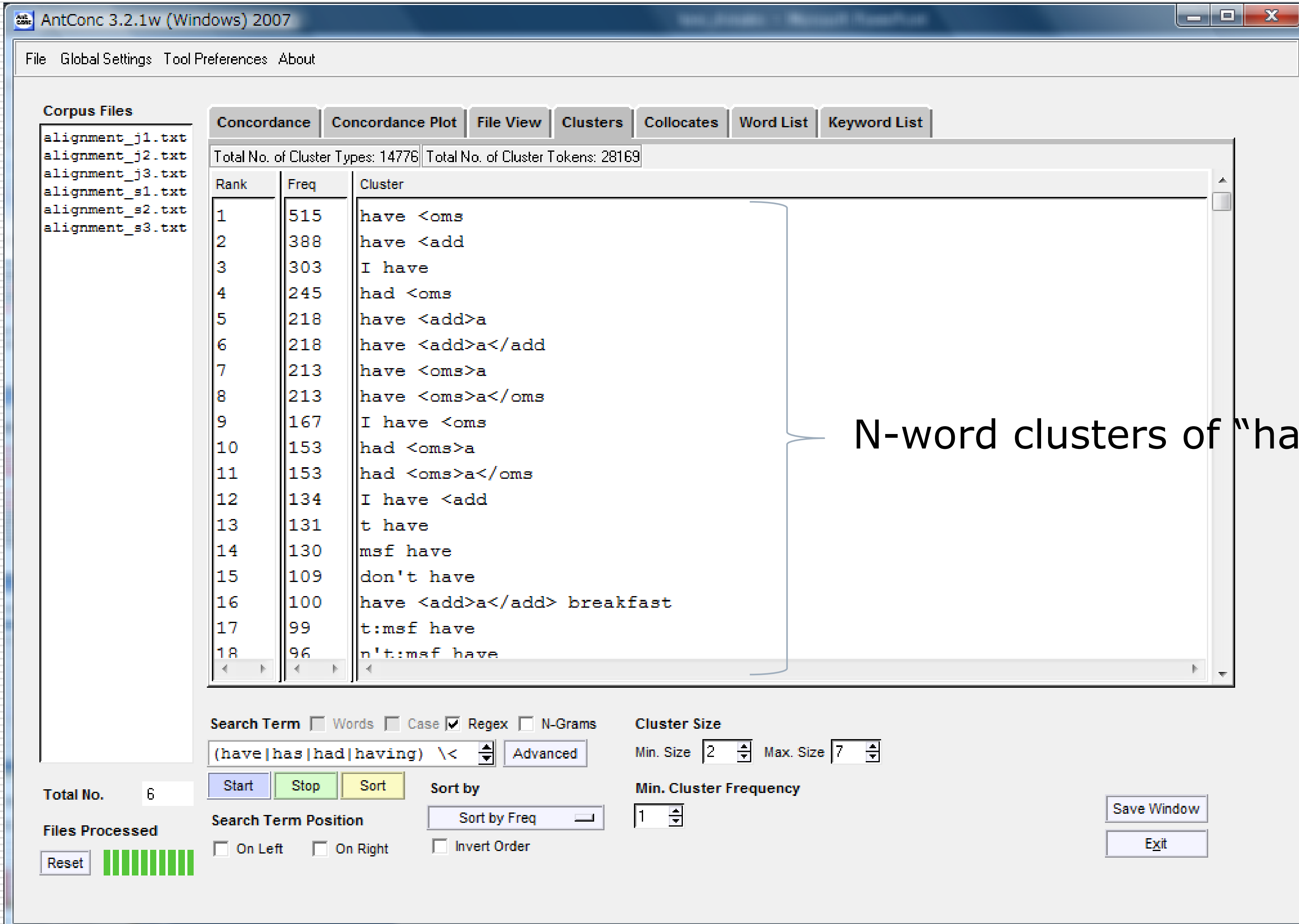


Use of modals

Marked omissions at the very beginning stages
Later, more use of modals lead to more addition errors



Errors related to ‘have’



N-word clusters of “have”

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

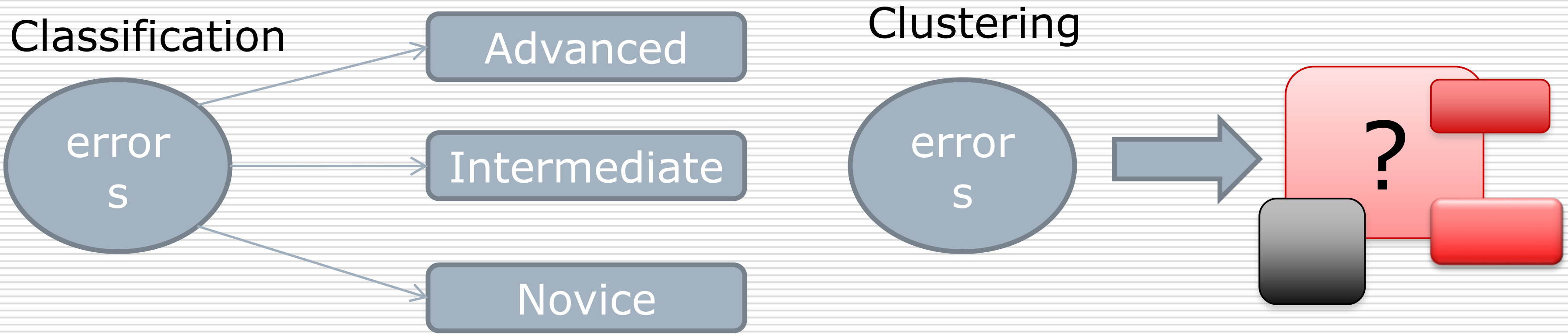
A: So I <oms>don't</oms> have <add>n't</add> time to eat breakfast

Supervised vs. unsupervised learning

Automatic extraction of error patterns from LC

Multivariate Analysis

Supervised Learning	Unsupervised Learning
Classification	Clustering

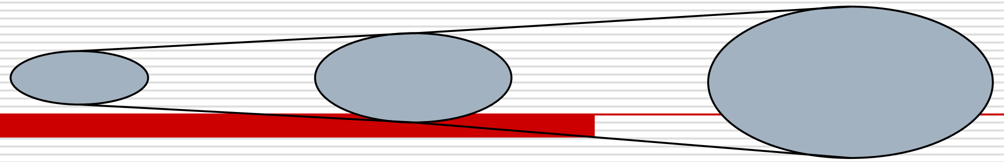


New project: ICCI

- 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