International Corpus of Japanese as a Second Language and Applications

Kumiko Sakoda

Hiroshima University

National Institute

for Japanese Language & Linguistics

INTERNATIONAL SYMPOSIUM ON DIVERSE APPROACHES TO SECOND LANGUAGE ACQUISITION : LEARNER CORPORA, EVALUATION AND BRAIN SCIENCES JUNE 1-2, 2019

Aims:

1. To introduce the large-scale learners' corpus of Japanese currently under construction

2. To analyze part of the data and discuss issues of grammar acquisition

Contents

1. Introduction

2. The learner corpus: **I-JAS**

(completion in 2020, next year!)
http://lsaj.ninjal.ac.jp/

3. A study: language use between different tasks (speaking vs. writing)

4. Conclusion

1. Introduction

1. Introduction
(learners' corpora in the past)

Table 1. List of Learners' corpora (spoken)

Corpus Name	Data	Learners' Native Languages	FS	Level Check
KY Corpus (Cross-Sectional)	90 (30min.)	Chinese, Korean, English	×	OPI
KAIWA-DB (Cross-Sectional)	339 (30min.)	Chinese, Korean, English Indonesian, others	○	OPI
BTSJ (Cross-Sectional)	294 dialogues 66 hours	Korean, Chinese, French	×	×

※ OPI : Oral Proficiency Interview (ACTFL)

Table 1. List of Learners' corpora (spoken) cont.

Corpus Name	Data	Learners' Native Languages	FS	Level Check
LARP (Longitudinal)	37 (20min) 3.5 years	Chinese	○	SPOT
KAIWA-DB (Longitudinal)	About 20 46 dialogues (30min.)	Tagalog, Korean, Chinese, Russian, Malay, Portuguese	×	OPI
C-JAS (Longitudinal)	6 47dialogues (60min) 3 years	Chinese, Korean	△	×

※ SPOT: Simple Proficiency Oriented Test

Low number of subjects

Corpus Name	Data	Corpus Name	Data
KY Corpus (Cross-Sectional)	90	LARP (Longitudinal)	37
KAIWA-DB (Cross-Sectional)	339	KAIWA-DB (Longitudinal)	About 20
BTSJ (Cross-Sectional)	294	C-JAS (Longitudinal)	6

Low number of countries

Corpus Name	Learners' Native Languages	Corpus Name	Learners' Native Languages
KY Corpus	Chinese, Korean, English	LARP	Chinese
KAIWA-DB	Chinese, Korean, English Indonesians, others	KAIWA-DB	Tagalog, Korean, Chinese, Russian, Malay, Portuguese
BTSJ	Korean, Chinese, French	C-JAS	Chinese, Korean

Face Sheet & Proficiency Level

Corpus Name	Face Sheet	Level Check	Corpus Name	Face Sheet	Level Check
KY Corpus	×	OPI	LARP	○	SPOT
KAIWA-DB	○	OPI	KAIWA-DB	×	OPI
BTSJ	×	×	C-JAS	△	×

Issues with Japanese learner corpora

1. Low number of learners
2. Most corpora contain data from English, Chinese or Korean native speakers; data for other languages is absent
3. Level of Japanese language proficiency is unclear
4. Background learner information is unavailable

2. The Learner corpus I-JAS

(International corpus of Japanese as a second language)

2． The learner corpus: I-JAS (under construction)

I-JAS

International corpus of Japanese As a Second language

【Aim】

To elucidate the effects on the acquisition process of different language environments, including differences in mother tongue

Characteristics of I-JAS

①Learners

・JFL Learners from 16 countries, speaking 12 native languages

・JSL Classroom/ Natural Setting

・Native Speakers

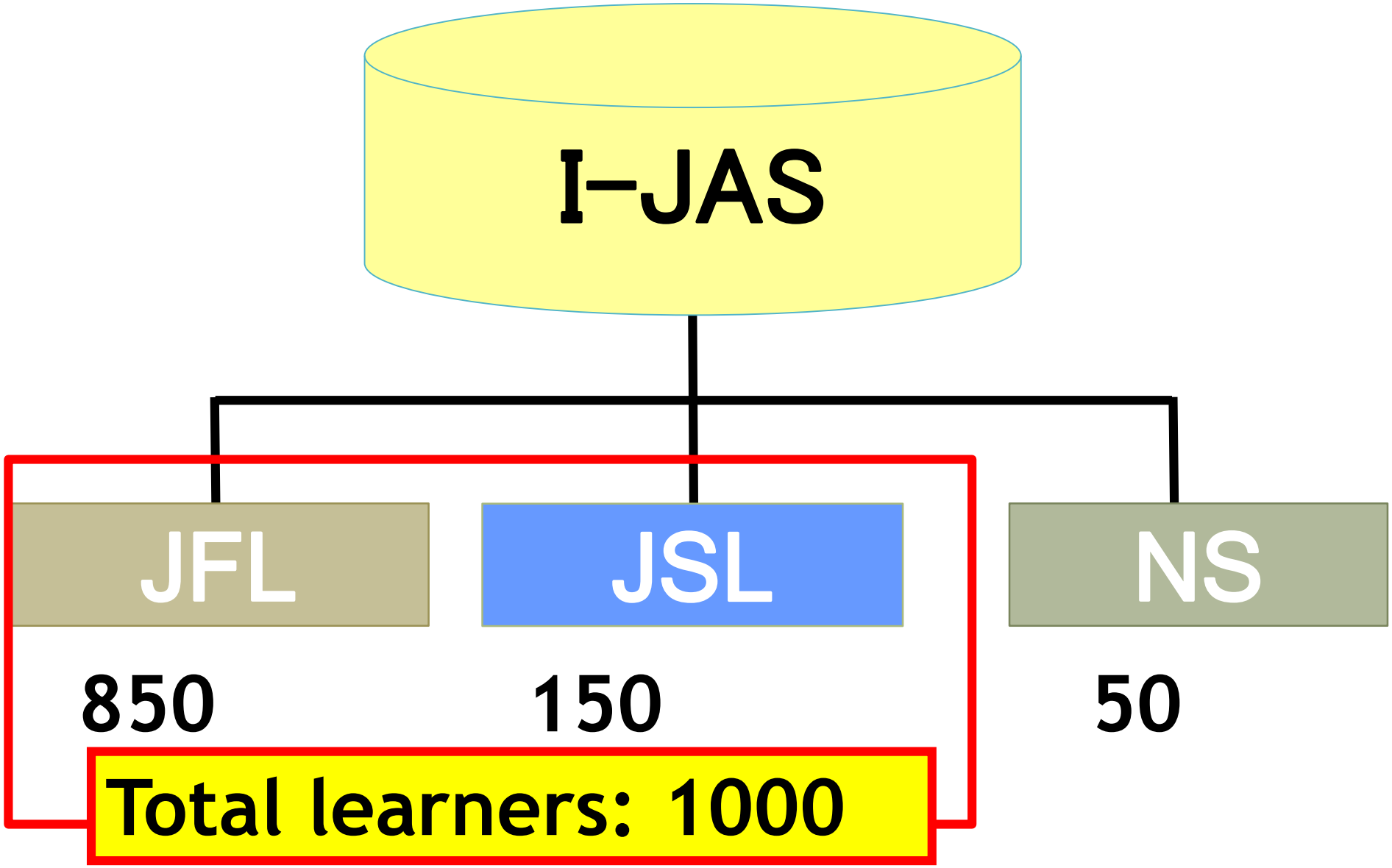
②Detailed background information

③Objective Japanese Proficiency Tests (2 types)

④A variety of tasks (6 types)

⑤Release of text and audio

Learners & Native Speakers



Learners of JFL

Chinese 200 learners	**German** 50 learners	**French** 50 learners
Korean 100 learners	**Turkish** 50 learners	**Spanish** 50 learners
English 100 learners	**Indonesian** 50 learners	**Russian** 50 learners
Thai 50 learners	**Vietnamese** 50 learners	**Hungarian** 50 learners

Learners of JSL & NS

Learners in Classroom Settings	Learners in Natural Settings	Native Speakers 20s 30s 40s
100	50	50

Corpus Design (speech)

1. Story Telling

Look at 4–5 pictures and tell the story

2. Dialogue (30 minutes)

Semi-structured interview

The previous day's schedule/interest in Japan/home town/childhood memories/future/opinions etc.

Story Telling
&
Story Writing

【PICNIC】

①

②

③

④

⑤

3. Role-play

"Refusal" and "request" tasks

4. Picture portrayal task

Look at and describe in Japanese a single image

5. Writing

Look at the pictures used in 1. and write the story

Xu(2000)

Corpus Design (writing)

1． Essay

“Our diets: fast food and home cooking”

(around 600 characters)

2． Email

Establish 3 scenarios,

then write emails

(request, refusal etc.)

Assessments of Japanese language proficiency

1. SPOT

(Simple Performance-Oriented Test)

Proficiency measured by testing aural comprehension

2. J-CAT

(Japanese Computerized Adaptive Test)

Computer-based proficiency test with automatic assessment

1st release of data
(Spring 2016)
225人

2nd release of data
(Spring 2017)
225人
450人

3rd release of data
(Spring 2018)
225人
660人

4th release of data
(Spring 2019)
215人
875人

last release of data
(Spring 2020)
175人
1050人

3. A study:
language use between different tasks
(speaking vs. writing)

Language use in different tasks

Thai Speaker (TH27)

◆**Speaking task** (error ×)

Inu wo tabete shimaimashita.

(We ate the dog.)

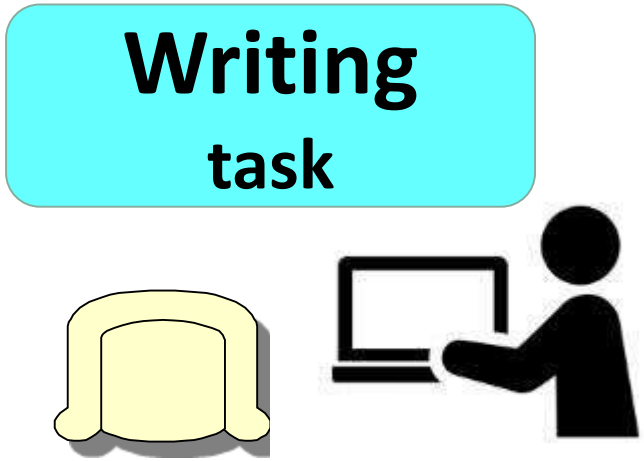
◆**Writing** (correct ○)

Inu ni taberarete shimaimashita.

(Sandwiches are eaten by the dog.)

■Research question

Is verb conjugation more accurate in the “writing” task?



■Learners (12 native languages)

Chinese, English, German, Indonesian, Hungarian, Russian, Spanish, French, Korean, Thai, Vietnamese, Turkish

Table 2. Results of the Proficiency Tests

	INDN.	ENG.	KRN.	SPN.	THAI	CHN.
J-CAT	209	210	211	189	211	211
SPOT	67	69	72	64	68	70
	GER.	TRK	HUNG	FRN	VIET	RSS
J-CAT	210	210	210	189	213	211
SPOT	70	69	69	66	65	70

Story-telling
(Picnic)

【SPEAKING】

The learner begins speaking.

- Interview Task
- Role Play
- Description Task

【WRITING】

The learner writes the story looking at the pictures.

①

②

③

④

⑤

Results of analysis (passive・～しまう) (Okuno 2015)

Table 3. Sentence Variations in Speaking and Writing

	INDN S	W	ENG S	W	KRN S	W	SPN S	W
V-rarete-shimatta	3	5	1	4	0	1	0	0
V-rareta	3	3	1	0	1	0	0	0
V-teshimatta	6	6	8	7	8	8	10	12
V-ta/ru	0	0	5	2	4	3	5	2
Others	3	1	0	2	2	3	0	1

Table 3. Sentence Variations in Speaking and Writing (Thai, Chinese, German & Turkish)

	THAI S	W	CHN S	W	GER S	W	TRK S	W
V-rarete-shimatta	0	4	2	1	1	3	2	0
V-rareta	4	3	8	11	1	0	1	2
V-teshimatta	8	8	0	0	10	10	5	3
V-ta/ru	3	0	2	2	3	1	6	7
Others	0	0	3	1	0	1	1	3

Results of analysis (passive・～しまう)

Table 3. Sentence Variations in Speaking and Writing (Hungarian, French, Vietnamese & Russian)

	NUNG S	W	FRN S	W	VIET S	W	RSS S	W
V-rarete-shimatta	0	1	1	0	2	5	1	1
V-rareta	1	1	0	4	2	1	2	5
V-teshimatta	11	10	8	7	6	4	5	4
V-ta/ru	2	2	4	4	4	2	6	3
Others	1	1	2	0	1	3	1	2

Change in forms used by the same learner

Table 4. Variations among Speaking and Writing by the same learner

	Speaking	Writing
THA 19	Tabe mashita	Tabe rarete shimaimshita
CHN 28	Tabe mashita	Tabe rare mashita
FRN 24	Tabe mashita	Tabete shimaimasita.
VTN 40	Tabete shimaimashita	Tabe rarete shimaimshita

- From Table 4

1. The written task data contains more instances of use of the passive (「rare」)form and the modal (「-teshimau」)form.

Supports Okuno (2015)

2. WE CAN SUPPOSE THE FOLLOWING ACQUISITION PROCESS

【PASSIVE】

TABE-MASHITA

TABE RARE MASHITA

Tabe rarete shimai mashita

TABE TA

TABETE SHIMATTA

Tabe rarete shimatta

【MODALITY】

Are the “writing” tasks more accurate grammatically than the “speaking” tasks?

Results of analysis (Sakoda 2019)

Table 5. Intransitive and Transitive Verbs

	Speaking Task (S)	Writing Task (W)	Intrans. / Trans. Verb
FRN 07	BK ni haitte shimatta ○	SD wo BK ni hairi (⇒irere) mashita ×	Hairu Ireru
VTN 51	Inu ga dashite (⇒dete) shimai mashita ×	Inu ga dete shimai mashita ○	Dashite Dete

SD = sandwich BK = basket

Results of analysis (Sakoda 2014)

Table 5. Intransitive and Transitive Verbs (cont'd.)

Student	Speaking Task (S)	Writing Task (W)	Intrans. / Trans. Verb
ENG 27	BK wo aita ato (⇒aketa ato) ×	BK wo aita (⇒aketa) tokoro ×	Aku Akeru
THAI 49	BK wo akuto (⇒akeru to) ×	BK wo akuto (⇒akeru to) ×	Aku Akeru

SD サンドイッチ　　BK バスケット

Results of analysis (passive・～しまう) (Okuno 2015)

Table 6. Error patterns of transitive/ intransitive verbs

	INDN	ENG	KRN	SPN	THAI	CHN
S○ W○	13	13	13	10	11	10
S○ W×	0	0	0	0	1	3
S× W○	0	0	0	1	0	0
S× W×	1	0	0	1	2	1
Others	1	2	2	3	1	1

Table 6. Error patterns of transitive/ intransitive verbs

	GER	TRK	NUNG	FRN	VIET	RSS
S○ W○	12	14	11	12	9	11
S○ W×	1	0	0	2	0	1
S× W○	1	0	0	1	1	1
S× W×	1	1	3	0	2	0
Others	0	0	1	0	3	2

From Table 6

There was no change observed in the use of intransitive and transitive verbs (both tasks showed the same usage trends)

For transitive-intransitive verb pairs, a tendency was observed to favor use of one or the other of the pair

Item learning ?

4 . Conclusions

What this study revealed:

(1) Using the same images to conduct "speaking" and "writing" tasks with the same learners, there were areas where differences were observed and those where none was observed.

Differences in the tasks (thinking time) may or may not have an effect

（2）There was a trend for passives and the「～te shimau」construction to be used when writing, even if they were not used in the speaking task

Tabe ta

Tabe mashi

Tabe rare ta

Tabeteshima

TABE RARETE SHIIMAI MASHITA

Learners have sufficient time to use correctly grammatical structures that they have studied

Syste m

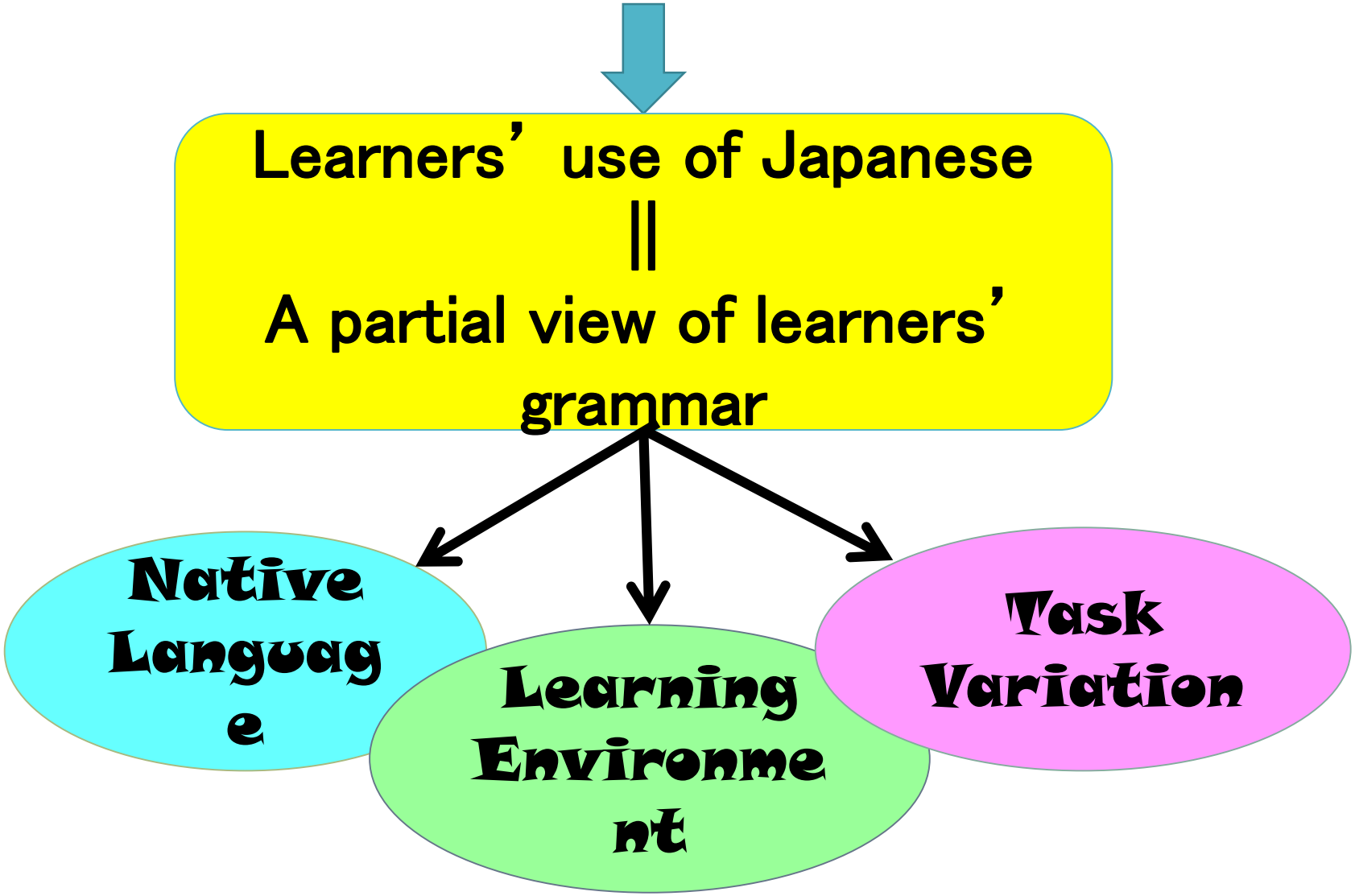
learning ?

(3) The trend is for there to be no change in the use of intransitive-transitive verb pairs between spoken and written language.

Intransitive and transitive verbs may be **being** processed as lexical rather than grammatical items

Item learning ?

What we can discover from learner corpora



References

1. Okuno, Y.(2015)「「Hanasu」kadai to「Kaku」kadai ni mirareru chuukangengo henni sei—Storybyousha Kadai ni okeru「Tabe rarete shimatta」bu wo taishoo ni—」Proceeding of NINJAL Workshop 2014, 20-23.
2. Sakoda, K. (2014) 「Kaki-kotoba to Hnashi koboba no Chigai-Gakushuusha Kopasu ni miru gengo unnyou」 -ICPLJ 2014
3. Sheu, Shiah-Pei. 2000. Shizen-hatsuwa ni okeru nihongo-gakushūsha niyoru “teiru” no shūtoku-kenkyū : OPI date no bunsekikekka kara *Nihongo Kyōiku* (104). 20-29.

Thank you for your attention.

Kumiko Sakoda

NINJAL