A Reconsideration of the Needed Sample Size in Learner Corpus Studies

A Reconsideration of the Needed Sample Size in Learner Corpus Studies

Shin'ichiro ISHIKAWA (Kobe University)

学習者コーパス研究における標本数の問題

石川　慎一郎（神戸大学）

Abstract

The number of samples collected in learner corpora is generally small in comparison to native speaker corpora, but the extent to which the limited sample size influences the reliability of learner corpus studies has not yet been wholly elucidated. Therefore, we extracted short writing pieces from the International Corpus of Japanese as a Second Language (I-JAS) and prepared text sets of different sizes (n = 10, n = 20, n = 30, n = 40, and n = 50) for Chinese and Korean learners of Japanese as well as Japanese native speakers. We then examined the difference ratios observed across five kinds of text sets with a focus on basic linguistic indices, such as the total number of tokens per texts, and frequencies of punctuation marks, nouns and verbs, and conjugation forms of verbs. Our analyses show that the influence of sample size is not as strong as generally expected, and that discussion of learners' L2 production with a relatively smaller corpus data could be rationalized to some extent.

1. Introduction

1.1 Sample Size of Learner Corpora

Recent developments in information technology have drastically expanded the size of native speaker corpora. In the 1960s, the Brown Corpus collected only 500 written text samples, while in the 1990s, the British National Corpus (BNC) collected more than 4,000 written text samples as well as speech samples by 124 volunteers. More recently, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) has collected 172,675 written text samples.

However, looking at learner corpora, we see a completely different picture. As learner corpora usually collect data from several learner groups with different L1 backgrounds, the number of samples for each group naturally tends to be smaller. The table below summarizes sample sizes of single learner groups in major English and Japanese learner corpora: the International Corpus of Learner English v1 (ICLE1) (Granger, Dagneaux, & Meunier, 2002), v2 (ICLE2) (Granger, Dagneaux, Meunier, & Paquot, 2009), International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013; Ishikawa, 2014), Louvain International Database of Spoken Interlanguage (LINDSEI) (Gilquin, De Cock, & Granger, 2010), Natane Learner Corpus (Nishina, Yagi, Hodošček, & Abekawa, 2014), Japanese Learner's Written Composition Corpus (JLWCC) (Lim, Lee, Miyaoka, Shibasaki, & Cho, 2013), International Corpus of Japanese as a Second Language (I-JAS) (Sakoda, Konishi, Sasaki, Suga, & Hosoi, 2016), and KY Corpus (KY) (Kamada, 2016).

Table 1. Sample Size of a Single Learner Group in Major English and Japanese Learner Corpora

Corpora	Target L2	Spoken/Written	# of learners	# of texts
ICLE1	English	W	N/A	246 – 392
ICLE2	English	W	N/A	243 – 982
ICNALE-Written	English	W	100 – 400	200 – 800
LINDSEI	English	S (OPI-like)	50 – 53	50 – 53
ICNALE-Spoken	English	S (Monologue)	50 – 150	200 – 600
Natane	Japanese	W	1 – 115	1 – 152
JLWCC	Japanese	W	N/A	144 – 160
I-JAS	Japanese	S (OPI-like) +W	50 – 200	50 – 200
KY	Japanese	S (OPI)	30	30

A learner group with a particular L1 background is usually subdivided into several proficiency groups. For example, learners are subdivided into A2, B11, B12, and B2+ levels in the ICNALE; and into Novice, Intermediate, Advanced, and Superior levels in the KY Corpus. This suggests that users of learner corpora often discuss a particular learner group on the basis of quite a small number of samples. For example, the number of Novice and Superior learners in the KY Corpus is five.

1.2 Needed Sample Size

This raises the question whether it is appropriate for us to study features of L2 learners' interlanguage use with such little data. In statistics, the needed sample size (n) is defined by (i) the size of population (N) (i.e., which size of population you have), (ii) the expected response rate (p) (i.e., which ratio of response you expect), (iii) the margin of errors (ME) (i.e., which level of errors you tolerate or which level of accuracy you need), and (iv) the z score (k) determined by the desired confidence interval (CI) (i.e., which level of confidence you need). The formula applicable to the finite population is as follows (Narita, 2001):

$$n = \frac{N}{\frac{N-1}{p(1-p)} \left(\frac{ME}{k}\right)^2 + 1}$$

P varies between 0 and 1. If 90% of the respondents say "yes" and the remaining 10% say "no" in some survey research, p is 0.9. And if 50% say "yes" and 50% say "no," p is 0.5. In the former case, the difference is large enough and we may allow a certain level of errors. While, in the latter case, errors should be closest to zero. As p is usually unknown beforehand, we often regard p as a constant of 0.5.

ME is chosen by researchers. When they require a greater level of accuracy, they will choose ±1% or ±5%; while when they do not need such a level of accuracy, they may choose ±10% or ±15%, for example.

CI is also chosen by researchers. When they need a greater level of confidence, they will choose 99% (k≒2.58) or 95% (k≒1.96); while when they do not, they may choose 90% (k≒1.65),

for example. Although 95% *CI* is usually chosen, this is not a fixed rule.

The needed number of samples (*n*) greatly varies according to *ME* and *CI* that we choose. The table below shows how many samples we need to collect when *N* is 10,000.

Table 2. Margin of Errors, Confidence Interval, and Needed Sample Size (*N* = 10,000)

	ME =±15%	*ME* =±10%	*ME* =±5%	*ME* =±1%
CI = 90% (*k*≒1.65)	30	68	264	4,036
CI = 95% (*k*≒1.96)	43	96	370	4,900
CI = 99% (*k*≒2.58)	74	164	623	6,240

If we need 90% accuracy (i.e., 10% *ME*) and 90% confidence, for example, the needed sample size is calculated as 68. This suggests that a certain level of accuracy and confidence can be realized even when dealing with a learner corpus of a relatively smaller size.

However, in learner corpus studies, the population, namely, the number of learners to be examined, can be much larger. How does this influence the needed number of samples? The table below shows the needed number of samples when *ME* is±10% and *CI* is 90%.

Table 3. Population Size and Needed Sample Size (*ME* = ±10% and *CI* = 90%)

N	*n*
10,000	68
100,000	68
1,000,000	68
10,000,000	68
100,000,000	68

The influence of *N* is quite small. For example, the numbers of Japanese learners in schools are reported to be 953,283 in China, 745,125 in Indonesia, 556,237 in Korea, 64,863 in Vietnam, and 33,234 in Malaysia (Japan Foundation, 2016). Such a difference, which seems to be substantial, should hardly influence the needed sample sizes. This can be explained by the law of large numbers (LLN) (i.e., errors become smaller according to the sample size) and the central limit theorem (CLT) (i.e., distribution becomes closer to normal according to the sample size).

As summarized above, statistics seems to support to some extent the validity of learner corpus studies using relatively smaller datasets, but how the sample size influences basic linguistic indices obtained from those datasets is not necessarily clear. The current study focuses on this matter.

2. Research Design

2.1 Aim and RQs

This study aims to empirically observe the relationship between the size of the sample and several quantitative indices obtained from it. Among various linguistic indices that have been used in corpus studies, we pay attention to (i) total number of tokens (text length), (ii) frequency of punctuation marks, (iii) frequency of nouns and verbs, and (iv) frequency of major verb conjugation forms. We

also discuss frequency of major POS types. Two research questions are discussed here.

RQ1. To what extent does the sample size influence the total number of tokens and the frequencies of punctuation marks, nouns and verbs, and major verb conjugation forms?

RQ2. How are different datasets clustered in terms of frequencies of major POS types?

2.2 Data

We analyze texts written by Chinese (CHN) and Korean (KOR) learners of Japanese, as well as Japanese native speakers (JPN) as a reference, all of which are taken from the I-JAS, 2nd version (released in June 2017). Currently, 50 participants' data have been released.

Table 4. The Number of Participants in CHN, KOR, and JPN Submodules of the I-JAS

	1st Release	2nd Release	Final Release
CHN ("CCM")	*n* = 15	*n* = 50 (15+35)	*n* = 200
KOR ("KKD"/"KKR")	*n* = 15	*n* = 50 (15+35)	*n* = 100
JPN ("JJJ")	*n* = 15	*n* = 50 (15+35)	*n* = 50

I-JAS collects varied types of learners' L2 Japanese productions: (1) story telling (two tasks based on serial pictures: "picnic" and "key"), (2) dialogue (free talk for thirty minutes), (3) role play (two tasks including "asking" and "refusal"), (4) picture description, and (5) story writing (two tasks based on the serial pictures used in the story telling), all of which were collected in face-to-face interviews, and in addition, (6) e-mail writing (three tasks) and (7) essay. The current study uses the "key" story writing data.

2.3 Method

First, we assigned serial numbers (#01 to 50) to the Chinese, Korean, and Japanese participants. Next, we made five kinds of datasets of different sizes for each of the these participant groups: *n* = 10 (#01 to #10), *n* = 20 (#01 to #20), *n* = 30 (#01 to #30), *n* = 40 (#01 to 40), and *n* = 50 (#01 to 50). We then processed all the texts with the Chasen morphological analyzer.

In order to examine the effect of sample size, we paid attention to the difference ratio (DR). DR is calculated by dividing the difference between the maximum and minimum values by the mean value. In the case of Chinese learners, for example, the average numbers of tokens per single texts are 122.9 (*n* = 10), 111.4 (*n* = 20), 118.1 (*n* = 30), 125.1 (*n* = 40), and 123.3 (*n* = 50). As the difference between the max and min values is 13.7 and the mean value is 120.8, DR is calculated as 11.3%. There are no fixed rules about how to interpret DR, but if DR is around 10% or lower than that, we could conclude that the sample size does not influence the result so strongly.

When discussing the total number of tokens, we calculated the number of morphemes occurring in each of the text sets. Punctuation marks and symbols are included, but spaces and end-of-sentence markers ("EOS") are excluded. Then, we calculates the average numbers of tokens per single texts.

Next, we investigated the number of full-stops (*kutens*) and commas (*toutens*) per 1,000 tokens. The comma/full-stop ratios, which reflect the degree of textual cohesion in texts, are discussed.

Also, we investigated the number of nouns and verbs per 1,000 tokens as well as noun/verb ratios. In the current study, we defined nouns and verbs in the broadest sense: nouns include common nouns, proper nouns, verbal nouns, bound nouns, and so on, and verbs include free verbs, bound verbs, suffix verbs, and so on. It is generally understood that a higher noun/verb ratio suggests an orientation toward the subject, transitivity, activeness, intention, action, and logic, while a lower ratio represents an orientation toward topic, intransitivity, inactiveness, unintentionality, event, and emotion (Ishikawa, 2015).

We then examined the number of continuative forms (*renyo-kei*), continuative-past forms (*renyo-"ta"-kei*), and basic forms (*kihon-kei*) of verbs per 1,000 tokens, all of which are the commonest conjugation forms and constitute more than 85% of all the verb forms.

Finally, concerning RQ2 (clustering), we conducted a hierarchical case cluster analysis on the frequency table with 15 datasets (three nationalities X five sample sizes) as cases and frequencies (per 1,000 tokens) of six major POS types (nouns, verbs, auxiliary verbs, free adjectives, particles, and conjunctions) as variables. For calculation of the distances and data agglomeration, Euclidean distance (standardized) and Ward's method are used.

3. Results and Discussions

3.1 RQ1 Difference Ratio

3.1.1 Total Number of Tokens per Texts

As a story writing task given in the I-JAS is based on the same serial pictures as a prompt, its length is expected to be largely the same. The results of the data analysis are shown below:

Table 5. The Average Number of Tokens per Single Texts

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	122.90	111.40	118.13	125.08	123.32	125.08	111.40	13.68	120.17	11.38
KOR	115.10	111.95	107.80	103.08	102.78	115.10	102.78	12.32	108.14	11.39
JPN	126.90	117.75	119.37	121.48	123.46	126.90	117.75	9.15	121.79	7.51

DR is 7.51% for JPN, and 11.38% and 11.39% for CHN and KOR, respectively. Although it is a bit larger than 10% for learners, we can generally conclude that the number of tokens is largely stable in spite of the sample size.

3.1.2 Punctuation

Punctuation marks are discussed frequently in corpus studies, for they are thought to reflect writers' unconscious preferences in writing. The results of the analysis are shown below:

Table 6. The Number of Full-stops

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	59.40	59.25	58.13	57.77	56.11	59.40	56.11	3.29	58.13	5.66
KOR	46.05	48.24	46.69	48.02	48.45	48.45	46.05	2.40	47.49	5.05
JPN	44.92	44.97	44.68	44.04	45.20	45.20	44.04	1.16	44.76	2.59

Table 7. The Number of Commas

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	65.91	61.49	66.03	64.36	65.20	66.03	61.49	4.54	64.60	7.03
KOR	43.44	33.05	40.20	37.84	39.50	43.44	33.05	10.39	38.81	26.77
JPN	44.13	44.54	43.28	44.25	45.52	45.52	43.28	2.24	44.34	5.05

Table 8. The Comma/Full-stop Ratio

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	1.11	1.04	1.14	1.11	1.16	1.16	1.04	0.12	1.11	11.17
KOR	0.94	0.69	0.86	0.79	0.82	0.94	0.69	0.26	0.82	31.54
JPN	0.98	0.99	0.97	1.00	1.01	1.01	0.97	0.04	0.99	3.88

The DRs for the number of commas and comma/full-stops ratio are exceptionally high for KOR (26.77% and 31.54%), but excluding KOR, DR falls between 3.88% and 11.17%.

3.1.3 Nouns/Verbs

The numbers of nouns and verbs are also discussed widely in corpus studies. The results of the analysis are as follows:

Table 9. The Number of Nouns

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	223.76	230.70	229.97	231.06	232.08	232.08	223.76	8.32	229.51	3.63
KOR	239.79	233.14	233.46	231.14	230.01	239.79	230.01	9.78	233.51	4.19
JPN	241.13	250.11	243.51	241.82	241.05	250.11	241.05	9.06	243.52	3.72

Table 10. The Number of Verbs

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	160.29	161.13	156.88	155.11	155.04	161.13	155.04	6.09	157.69	3.86
KOR	173.76	175.97	177.18	177.06	175.91	177.18	173.76	3.42	175.98	1.94
JPN	186.76	187.15	190.17	192.43	189.21	192.43	186.76	5.67	189.14	3.00

Table 11. The Noun/Verb Ratio

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	1.40	1.43	1.47	1.49	1.50	1.50	1.40	0.10	1.46	6.93
KOR	1.38	1.32	1.32	1.31	1.31	1.38	1.31	0.07	1.33	5.62
JPN	1.29	1.34	1.28	1.26	1.27	1.34	1.26	0.08	1.29	6.19

Concerning the frequencies of nouns and verbs, DR is quite small and falls between 1.94% and 6.93%. The influence of the sample size is hardly seen here.

3.1.4 Conjugation Forms of Verbs

Unlike the other indices, the numbers of conjugation forms of verbs are expected to be less stable, for many textbooks give priority to covering all conjugation forms and do not necessarily focus on making learners understand the usage of frequent verb forms (Noda, 2006). Thus, novice learners are likely to use conjugation forms in a deviant way, which leads to a greater variation between individual learners. The results of our investigation are shown below:

Table 12. The Number of Basic Forms

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	16.27	15.71	14.95	14.79	15.89	16.27	14.79	1.48	15.52	9.53
KOR	21.72	20.54	20.72	21.10	20.63	21.72	20.54	1.18	20.94	5.63
JPN	27.58	23.98	22.34	20.58	19.93	27.58	19.93	7.65	22.88	33.43

Table 13. The Number of Continuative-past Forms

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	32.55	33.66	29.63	28.78	27.41	33.66	27.41	6.25	30.41	20.56
KOR	35.62	31.71	33.09	31.53	31.33	35.62	31.33	4.29	32.66	13.14
JPN	33.10	31.26	33.79	34.78	35.15	35.15	31.26	3.89	33.62	11.57

Table 14. The Number of Continuative Forms

	n=10	*n*=20	*n*=30	*n*=40	*n*=50	Max	Min	Dif	Av	DR
CHN	95.20	90.22	92.83	92.34	92.77	95.20	90.22	4.98	92.67	5.37
KOR	98.18	100.94	99.88	101.63	99.82	101.63	98.18	3.45	100.09	3.45
JPN	108.75	113.06	115.33	116.69	114.53	116.69	108.75	7.94	113.67	6.99

It is true that DR is quite high in terms of the numbers of basic forms for JPN and continuative-past forms for CHN, but excluding these, DR falls between 3.45% and 13.14%. The effect of sample size may not be as substantial as generally expected even with a relatively unstable linguistic index like frequencies of conjugation forms. Additionally, it should be noted that the DR for native speakers, which is usually much lower than that for learners, can be higher in some cases.

3.2 RQ2 Clustering

A cluster analysis reveals how different subgroups are clustered together. Here we have two factors that potentially influence the result of clustering, namely, L1 and sample size. The learner corpus data is usually classified according to participants' L1. If so, we will have a CHN cluster, a KOR cluster, and a JPN cluster. However, if the sample size strongly influences the frequencies of major POS types, we might have mixed subgroups with different L1s. Theoretically, three patterns can be presupposed: (1) a CHN cluster, a KOR cluster, and a JPN cluster are neatly classified (when the sample size effect is small), (2) a JPN cluster, which is expected to show a higher level of stability and consistency in POS distribution, is observed, while CHN and KOR data are mixed up and do not form

neat clusters (when the sample size effect is middle), and (3) all the data are mixed up and no clear clustering can be seen (when the sample size effect is large).

Based on cluster analysis, we obtained the tree diagram below:

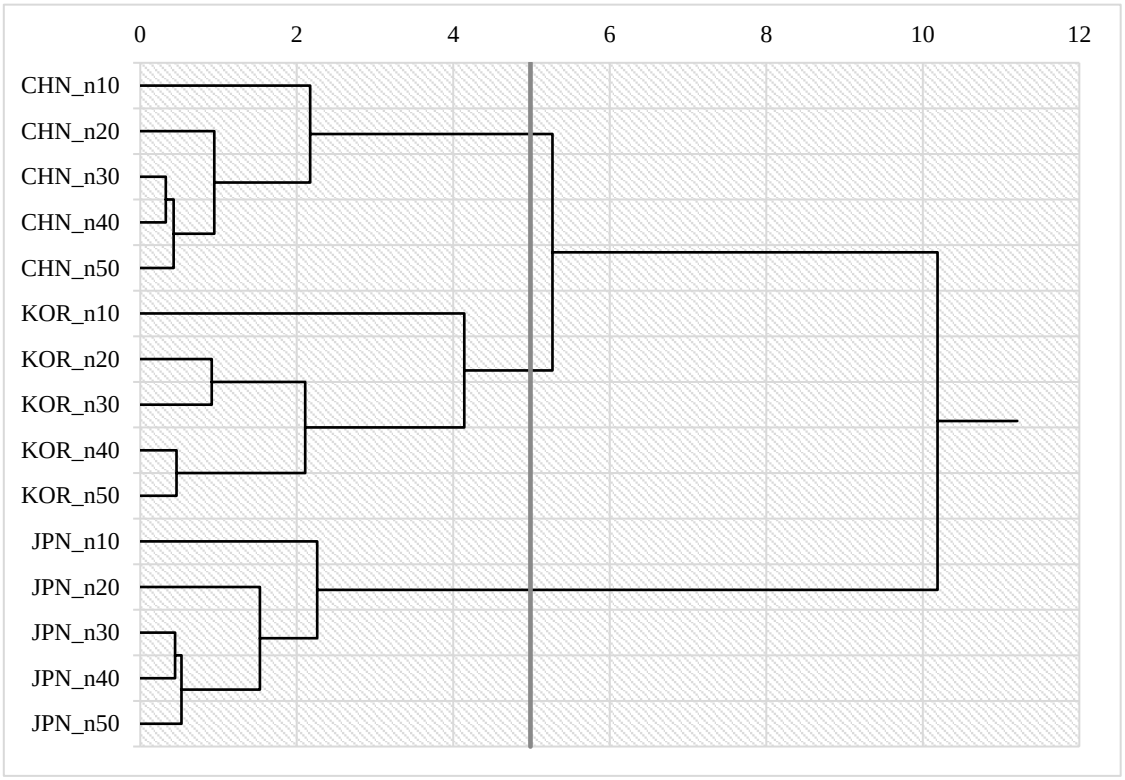

Fig. 1. Tree Diagram Obtained From a Hierarchical Case Cluster Analysis

The tree diagram shows that the data are neatly classified into three independent clusters based on participants' L1, which seems to support hypothesis (1). The sample size, though it influences the results, may not be a decisive factor, at least when discussing basic linguistic features such as POS frequencies. Another finding here is that $n = 50$ and $n = 40$ samples (and also $n = 30$ samples for CHN and JPN) are clustered together at a very early stage, suggesting that a variation between individual participants begins to converge when the sample size reaches 30 or 40.

4. Conclusion

The current study investigated how the sample size influences the basic linguistic features obtained from learner corpora. Concerning RQ1 (the difference ratio), we found that the difference ratio, calculated by dividing the difference between the maximum and the minimum values by the mean values, ranged across five kinds of text sets of different sizes from 7.51–11.38% for the total number of tokens per texts, 3.88–31.54% for the frequencies of punctuation, 1.94–6.93% for the frequencies of nouns and verbs, and 3.45–33.43% for the frequencies of major conjugation forms of verbs. Although the difference ratio surpassed 30% in a few cases, it proved to be under approximately 10% level in most cases. Also, concerning RQ2 (clustering), we confirmed that text sets were neatly classified according to participants' L1 in spite of the differences of sample size and that variation

began to converge in the samples of $n = 30$ or $n = 40$. These findings might rationalize to some extent the validity and reliability of the analysis of learners' L2 production using a relatively small corpus data.

However, there remain several limitations in the current study. First, it is debatable whether the statistical methods to determine the needed sample size can be applied to corpus studies. Second, we examined only a few basic linguistic indices in this study, and whether the same trend can be seen with other varied indices remains unclear. Third, as the available data was limited, we could not conduct a random sampling in preparing five kinds of datasets, which may make the difference rates across datasets seem lower than they really are. Also, even if our studies with smaller data sets can be rationalized, this should never negate the importance of observing learners' L2 use with larger data sets.

Bibliography

Gilquin, G., De Cock, S., & Granger, S. (2010). *Louvain international database of spoken English interlanguage*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., & Meunier, F. (Eds.). (2002). *International corpus of learner English*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English. Version 2*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian Learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University.

Ishikawa, S. (2015). Noun/verb ratio in L1 Japanese, L1 English, and L2 English: A corpus-based study. *Proceedings of the Second International Conference on Language, Education, Humanities & Innovation 2015*, 134-145.

Japan Foundation (2016). 2015 nendo kaigai nihongo kyoiku kikan chosa kekka: Sokuhochi. Tokyo: Japan Foundation. [Result of the survey on overseas Japanese language teaching institutes for the academic year 2015]

Kamada, O. (2016). KY kopasu to nihongo kyoiku kenkyu. *Nihongo Kyoiku*, *130*, 42-51. [KY Corpus and studies of Japanese language teaching]

Lim, H., Lee, J., Miyaoka, Y., Shibasaki, H., & Cho, G. (2013). Gengo shori gijutsu o riyo shita nihongo gakushusha sakubun kopasu no kaihatsu. *Nihon Bunka Gakuho*, *56*. [Development of the Japanese learner's written composition corpus based on language processing techniques]

Narita, S. (2001). Hyohon wa ikutsu atsumetara yoika. Retrieved from http://www.ceser.hyogo-u.ac.jp/naritas/spss/sample_size/sample_size.htm [How many samples should we collect?]

Nishina, K., Yagi, Y., Hodošček, B., & Abekawa, T. (2014). Construction of a learner corpus for Japanese language learners: Natane and Nutmeg. *Acta Linguistica Asiatica*, *4*(2), 37-51.

Noda, H. (2006). Komyunikeshon no tame no nihongo kyoiku bumpo. *Nihongo Kyoiku Tsushin*, *54*, 14-15. [Pedagogical Japanese grammar for communication]

Sakoda, K., Konishi, M., Sasaki, A., Suga, W., & Hosoi, Y. (2016). Tagengo bogo no nihongo gakushusha odan kopasu. *NINJAL Project Review*, *6*(3), 93-110. [On the international corpus of Japanese as a second language]