ISSN 1881−445X

日本語用論学会

■

第25回
大会発表論文集

第18号

Proceedings of the 25th Conference of the Pragmatics Society of Japan

2022年11月26日（土）・27日（日）

■

於　京都大学　吉田キャンパス・オンライン開催

PSJ

The Pragmatics Society of Japan
日本語用論学会
2022

Historical Pragmatics using State-Space Models[1]

Akitaka Yamada
a.yamada.hmt@osaka-u.ac.jp

Abstract. In Historical Pragmatics, we cannot easily rely on contemporary speakers' introspections. Therefore, quantitative examination of frequency tables assumes particular importance. A commonly used statistical technique is Chi-square analysis. However, this analysis has a number of limitations that make it unsuitable for investigating chronological time series data. Moreover, too much dependency on Chi-square analysis may result in biased and/or misleading interpretations. To overcome these shortcomings, this paper recommends the use of State-Space Models, more specifically, the Bayesian implementation of Dynamic Generalized Mixed-Effects Models (DGMM). The advantages of these models are demonstrated in two case studies that examine the variations found in Japanese honorific constructions.

Keywords: Chi-square analysis, State-Space Model, Historical Pragmatics, honorifics

1. Introduction: methodological issues in Historical Pragmatics

Historical Pragmatics is the study of language use in the past, and its development over time — although there are differences in definition between the Anglo-American tradition and the Continental European style (Goossen 1995; Jucker 2008; Taavitsainen and Jucker 2010). From a methodological perspective, one controversy is whether to study a language by relying solely on spoken data. In recent literature, however, the clear-cut dichotomy between spoken and written language has been challenged, and instead their continuity has been emphasized (Jucker 2008; Taavitsainen and Jucker 2010). Under this reassessment, each piece of writing is now seen as a mixture of different modes of language, thus laying a theoretical foundation for justifying the use of written texts in Historical Pragmatics.

However, this does not mean that all methodological issues are settled. A lingering problem for Historical Pragmatics which, in my view, has not received enough attention is the choice of quantitative methods. Of course, quantitative studies are not the only research method, and it is not my intent to promote one method over another. But in most cases, practitioners of Historical Pragmatics have no choice but to develop their arguments by relying on the frequency of constructions of their interest; digital corpus archives make this task easy. No one would deny that statistical methods are useful for fine-grained investigations. Nonetheless, when compared to other fields of humanities or social sciences (such as economics, psychology, and sociology), advanced statistical methods have not been as widely used. As a result, simplistic statistical models are blindly utilized to the extent that interpretating the data becomes biased or even misleading. One such practice is the use of Chi-square analysis.

The purpose of this present study is, therefore, to introduce an underused and advanced

[1] There was a typo in the title included in the program: *Space-State → State-Space.

	1901-1950	1951-2000
Construction A	20	35
Construction B	44	32

Table 1 A hypothetical classification table for Chi-square analysis.

statistical model, the State-Space Model, and to embed Historical Pragmatics into the larger context of Digital Humanities (Blei and Lafferty 2006, McCart 2014). After reviewing the limitations of descriptive and inferential statistics using Chi-square analysis, this paper presents the basics of the State-Space Model, and demonstrates how the analysis is applied to chronological data by taking two recent studies of honorifics for our primary examples.

2. Common practices in the literature

2.1. Descriptive statistics

A departure-point of a corpus/philological study is to create a table summarizing the frequency of the observed constructions (see for example Table 1). It is possible to develop an analysis that directly interprets the numbers in the table, but this line of descriptive approach has several disadvantages. First, the interpretation may depend too much on the researcher's subjective impression. Second, the table is just a sample of a larger population, the understanding of which is, in most cases, the ultimate goal of linguistic inquiry.

2.2. Inferential statistics: Chi-square analysis

To overcome these problems, inferential statistics has been developed, which allows us to make an inference about the structure of the population in a less subjective manner. For example, Chi-square analysis can be applied to the data in Table 1 ($\chi^2 = 5.0897$, df = 1, p-value = 0.02407). Under the commonly assumed threshold of $\alpha = 0.01$, it would be concluded that the null hypothesis that there is no difference between the two time periods is maintained. In such a setting, one cannot argue for the presence of a language change. Note that, as an elaborate modification, Fisher's exact test is utilized for smaller sample sizes (aka Collostructional Analysis, Gries and Stefanowitsch 2004; Stefanowitsch 2013), but it is essentially the same as Chi-square analysis in that it applies to data with a discrete classification table.

As common as these tests are, they are not suitable for diachronic corpus studies for at least the following reasons. First, for Chi-square analysis, we need to categorize the otherwise 'continuous' time variables into an arbitrary 'discrete' set of time periods. In Table 1, the samples are classified as either occurring before 1951 or after 1950, but this cut-off point is completely arbitrary, and the results differ under different cut-off points. For example, by using two more cut-off points, one could create the results seen in Table 2. Then, when Chi-square analysis is applied to these numbers, a very small p-value is obtained ($\chi^2 = 16.999$, df = 3, p-value = 0.0007), so unlike in our previous table, we could conclude that there was a change in language use, despite the fact that the data itself remains unchanged.

Second, in Chi-square analysis, independent fixed effects variables are not incorporated. The lack of explicit manipulation of independent/confounding variables can result in a biased interpretation. For example, suppose that Construction A is favored in the past tense

	1901-1925	1926-1950	1951-1975	1976-2000
Construction A	10	10	10	25
Construction B	22	22	22	10

Table 2 Using different cut-off points for the data in Table 1.

	1901-1950	1951-2000
Construction A	200	350
Construction B	440	320

Table 3 A classification table with a larger sample size.

throughout the 20th century, and the relatively large number of uses of Construction A during 1976–2000 in Table 2 is, in fact, not attributed to language change, but to the fact that there are many more past tense constructions collected than from the other three time periods. Naïve classification tables such as Table 2 fail to differentiate the effect of such confounding variables from the real chronological change. Thus, we need to explicitly incorporate intra/extralinguistic variables, whether they are fixed-effects or random-effects factors.

Finally, a large sample size results in a very small p-value, regardless of there being a substantial difference in population. Unlike in traditional experiments, where human subjects are recruited and the sample size is often smaller than in corpus linguistics, the data size used in corpus linguistics is quite large and thus tends to give a smaller p-value, which in turn makes a Type I error more likely. For example, the ratio in Table 3 is exactly the same as in Table 1. Nonetheless, due to the large sample size, the null hypothesis is rejected with a small p-value ($\chi^2 = 58.342$, df = 1, p-value = 2.203e-14).

3. State-Space Model

Given the discussion so far, an ideal model for statistical analysis of diachronic language change should meet the following requirements:

(1) Desiderata
  a. Track down the 'continuous' language shift/change as clearly as possible.
  b. Incorporate fixed-effects/random-effects variables — intra-/extralinguistic factors.
  c. Deal with the p-value issue by using a large sample size.

The Bayesian estimation of elaborate Dynamic Generalized Mixed-Effects Models (DGMM), as described below (Hagiwara 2021), is demonstrated to have these desired properties, and hence to be superior to Chi-square analysis in all respects.

3.1. State-Space Model

In the State-Space Model, observed values at time *t* are assumed to be generated on the basis of the latent state at a given time, which is related only to the state at the previous time point. For example, suppose that the *i*-th observed value at time *t*, $y_i^{(t)}$, is a binary outcome variable,

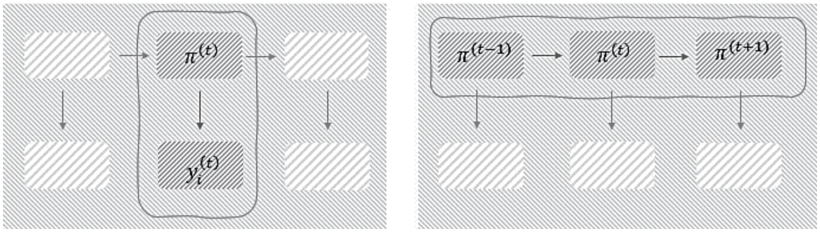

Figure 1 State-Space Model: (Left) Observation equation; (Right) State equation.

representing whether the meaning in question is expressed by Construction A ($y_i^{(t)} = 0$) or Construction B ($y_i^{(t)} = 1$). It is assumed that the probability of using Construction B is determined by the latent state $\pi^{(t)}$ at the given time point. If $\pi^{(t)}$ is large, then Construction B is likely to be pronounced, and if it is small, then Construction A is more likely. Researchers cannot directly observe every value of $\pi^{(t)}$, and it is a hypothesized construct; for this reason, it is called the LATENT STATE. The left panel in Figure 1 captures this relation between $y_i^{(t)}$ and $\pi^{(t)}$, and the mathematical formula describing this relation is called the OBSERVATION EQUATION. In our case, the equation is expressed as follows:

(2) $y_i^{(t)} \sim \mathrm{Bern}\left(\pi^{(t)}\right)$

The latent state is, however, not assumed to be stable across time; if it becomes larger, then Construction B becomes more popular. The right panel in Figure 1 represents the change of $\pi^{(t)}$, and the formula describing this chronological relation is called the STATE EQUATION. While there is a degree of freedom regarding how we postulate the relation, a commonly used model is the one with the random-walk structure, as shown in (3). The main objective of the State-Space Model is to estimate the magnitude of each latent state, so we can track down the trend in time series data.

(3) $\pi^{(t)} \sim N\left(\pi^{(t-1)}, \sigma_w^2\right)$

3.2. Dynamic Generalized Mixed-Effects Model (DGMM)

The aforementioned State-Space Model can be extended by explicitly incorporating intra- and extralinguistic factors. For example, irrespective of the year, Construction B may be preferred when it is used in the past tense; some predicates are more likely to be used in Construction B, or Construction B is more easily produced in some genres (sociolinguistic environments). Just as in the practice of Variation Theory (Cedergren and Sankoff 1974), the presence or absence of certain linguistic features is coded as a fixed-effect variable, and the idiosyncratic property of open-class elements (such as verbs and genres) as a random variable (Yamada 2022a, b). For example, by taking the i-th observation of the j-th genre at given t, the value of $\pi_{ij}^{(t)}$ can then be modeled as a linear combination of the intercept $\gamma_{00}^{(t)}$, the random variable u_{0j}, and the

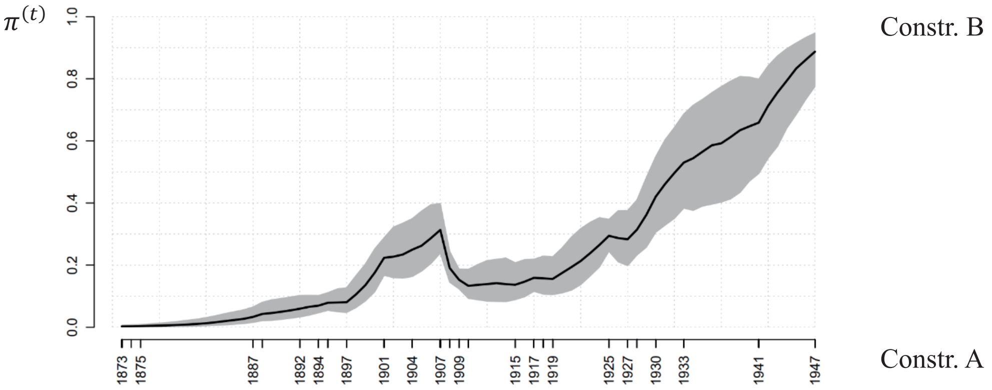

Figure 2 Posterior distributions for $\pi^{(t)}$.

fixed-effects term βx_i (aka the Dynamic Generalized Mixed-Effects Model or DGMM):

(4) $\pi_{ij}^{(t)} = \text{inv\_logit}\left(\gamma_{00}^{(t)} + u_{0j} + \beta x_i\right);\ u_{0j} \sim N(0, \sigma_{00}^2)$

In a recent development in computation statistics, estimation becomes fairly easy within the Bayesian framework; the examples introduced in Section 4 are estimated by Stan on R (Gelman et al. 2013). For example, Figure 2 shows what the results of DGMM looks like. Based on the solid line (the posterior median), and the gray-shaded area (the credible intervals), we can interpret how the probability of Construction B changes. Notice that in DGMM, we do not segment the time points into a set of arbitrary time ranges, thus circumventing the problem of Chi-square analysis in (1)a. As more clearly shown in Section 4, the posterior distributions for fixed- and random-effects variables are also calculated (= (1)b), and since the shaded area captures the uncertainty, researchers can make an inference without resorting to the notorious p-value (= (1)c), as is done in Frequentist statistics. In this way, the three desiderata in (1) are clearly satisfied. DGMM is superior for diachronic linguistic data in all respects.

4. Case studies

4.1. Case 1: Development of the use of *des*- in the canonical adjective construction

The honorific allocutivity (i.e., the addressee-honorification system) underwent a significant change in the 20th century (Kawaguchi 2014); Construction B in (5) became gradually more popular, making the old variant (Construction A) less common.

(5) a. Construction A: [canonical adj. (*i*-adj.)] + *gozai-mas* ‘ADJ-HON$_{\text{U}}$-HON$_{\text{A}}$’
b. Construction B: [canonical adj. (*i*-adj.)] + *des* ‘ADJ-HON$_{\text{A}}$’

Yamada (2022a) applied DGMM to the data in CHJ (the Corpus of Historical Japanese), assuming the structure in (6) for the population, and revealed how the change was affected by several extra-/intralinguistic factors.

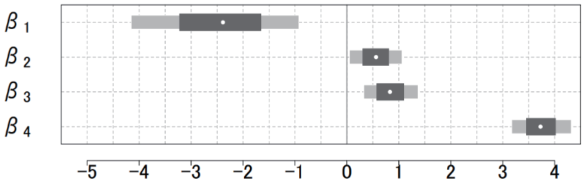

Figure 3 Posterior distributions for the fixed effects.

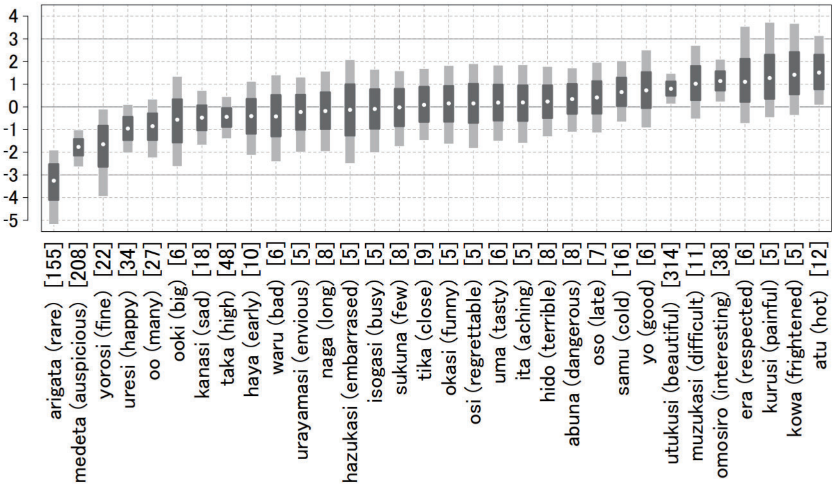

Figure 4 Posterior distributions for the random effects (i.e., the canonical adjective).

(6) $y_{ij}^{(t)} \sim \mathrm{Bern}\left(\mathrm{inv\_logit}\left(\beta_0^{(t)} + \beta_1 x_{1i}^{(t)} + \beta_2 x_{2i}^{(t)} + \beta_3 x_{3i}^{(t)} + \beta_4 x_{4i}^{(t)} + u_{0j}\right)\right)$

$$u_{0j} \sim N(0, \tau^2); \beta_0^{(t)} \sim N\left(\beta_0^{(t-1)}, \sigma_\zeta^2\right)$$

The estimated posterior distributions for $\pi^{(t)} = \mathrm{inv\_\ logit}\left(\beta_0^{(t)}\right)$ are shown in Figure 2, and those of the fixed- and random-effects are given in Figures 3 and 4, respectively.

4.2. Case 2: Diachronic alternation between *sase-te kure* and *sase-te moraw*

When a speaker describes an event, the execution of which is permitted by an authoritative and honorable person, one of the applicative constructions in (7) is chosen (Yamada 2022b).

(7) a. Construction A: -*sase-te* *kure (kudasar)* '-CAUS-CV APPL (APPL.HON)'
b. Construction B: -*sase-te* *moraw (itadak)* '-CAUS-CV APPL (APPL.HON)'

To reveal the historical change of these constructions, Shiina (2021) conducted a corpus study using BCCWJ and Aozora Bunko. Applying Chi-square analysis, she concluded that the non-honorific use of Construction A (-*sase-te kure*), and the honorific use of Construction B became popular in the 20th century.

However, Yamada (2022b), who re-analyzed the Aozora Bunko corpus (the same data examined by Shiina 2021) points out that the aforementioned generalization is not easy to

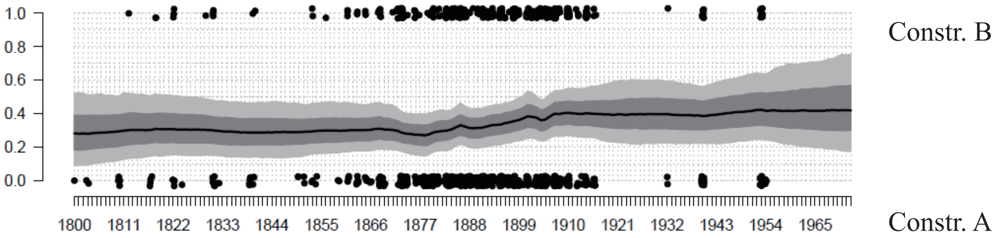

Figure 5 Estimated probability of *-sase-te moraw* over *-sase-te kure* in Aozora Bunko

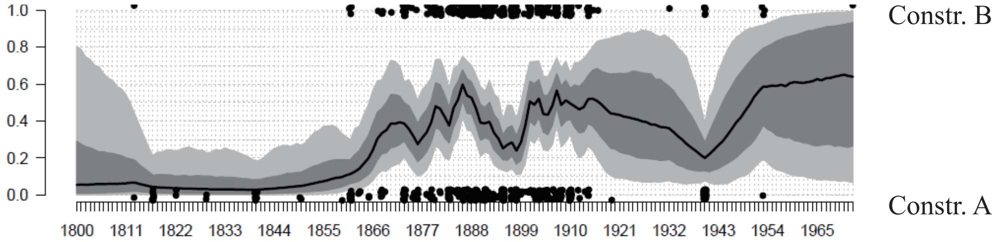

Figure 6 Estimated probability of *-sase-te itadak* over *-sase-te kudasar* in Aozora Bunko

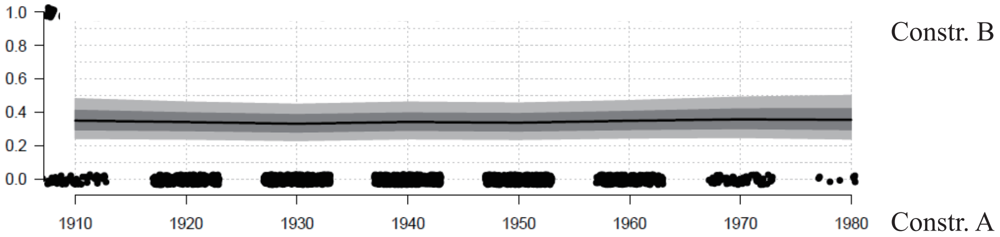

Figure 7 Estimated probability of *-sase-te moraw* over *-sase-te kure* in BCCWJ

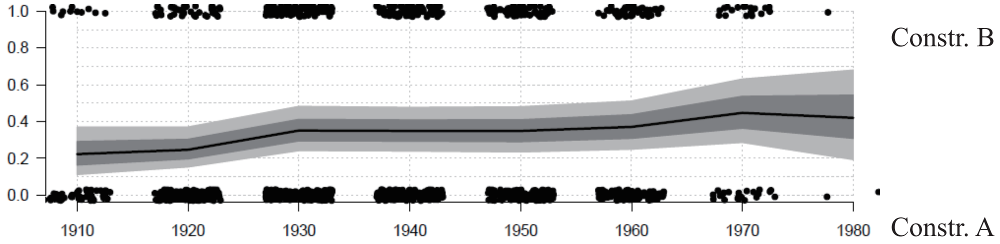

Figure 8 Estimated probability of *-sase-te itadak* over *-sase-te kudasar* in BCCWJ

maintain with the estimated results of the parameters of the following Time Series model assumed for the population (Figures 5 and 6); the subscripts *i* and *j* represent the *i*-th observation of the *j*-th verb and $\gamma_{00}^{(t)}$ is the intercept for the probability of producing Construction B at *t*, which has a random walk structure with variance σ_w^2; u_{0j} represents the uniqueness of the verb *j*, and β_1 is the coefficient for the honorification.

(8) $y^{(t)} \sim \text{Bern}\left(\text{inv\_logit}\left(\gamma_{00}^{(t)} + u_{0j} + \beta_1 x_{ij}\right)\right);\ u_{0j} \sim N(0, \sigma_{00}^2);\ \gamma_{00}^{(t)} \sim N\left(\gamma_{00}^{(t-1)}, \sigma_w^2\right)$

Additionally, Yamada (2022b) examined the data from BCCWJ, and suggested (i) that the ratio between *-sase-te kure* and *-sase-te moraw* remains stable irrespective of the author's birth year; and that (ii) while the use of *-sase-te itadak* came into use in the 19th century — for the simple reason that this construction was never utilized in any earlier period — it did not become popular enough to outnumber the competing construction (*-sase-te kudasar*) (Figures 7 and 8). In this way, Time Series Analysis enables us to make a finer-grained analysis of language change, which would not be so easy to detect with Chi-square analysis.

References

Blei, D. M. & John D. L. 2006. Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*, 113-120.

Cedergren, H. J. & Sankoff, D. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50(2). 333–355.

Gelman, A., Carlin, J. B.. Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. 2013. *Bayesian Data Analysis* [3rd edition]. London: CRC Press.

Goossens, L. 1995. Historical Linguistics. *Handbook of Pragmatics: Manual*, ed. by J. Verschueren, J.-O. Östman & J. Blommaert, 323–9. Amsterdam: John Benjamins.

Gries, S. Th. & A. Stefanowitsch. 2004. Extending Collostructional Analysis: a Corpus-Based Perspective on 'Alternations.' *International Journal of Corpus Linguistics* 9(1), 97-129.

Jucker, A. H. 2008. Historical Pragmatics. *Language and Linguistics Compass* 2(5), 894-906.

Hagiwara, J. 2021. *Time Series Analysis for the State-Space Model with R/Stan*. Singapore: Springer.

Kawaguchi, R. 2014. *Teineitai Hiteikei no Barieeshon ni Kansuru Kenkyuu [A Study of Variations among Negative Polite Forms]*. Tokyo: Kuroshio Publishers.

McCart, T. M. 2014. *A Statistical Analysis of Witchcraft Accusations in Colonial America: A Time Series Count Data Analysis*. Ph. D. Thesis, Youngstown State University.

Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Shiina, M. 2021. *"Saseteitadaku" no Goyooron: Hito wa Naze Tukaitaku Naru no ka [The Pragamtics of "Saseteitadaku": Why do People Want to Use it?]*. Tokyo: Hituzi Syobo.

Stefanowitsch, A. 2013. Collostructional Analysis. In T. Hoffmann, and G. Trousdale (eds), *The Oxford Handbook of Construction Grammar*. 290-306. Oxford: Oxford University Press.

Taavitsainen, I. & A. H. Jucker. 2010. *Historical Pragmatics*. Berlin: Mouton de Gruyter.

Yamada, A. 2022a. Constructionalization of the Japanese Addressee-Honorification System. The 23rd Annual Meeting of the Japanese Cognitive Linguistics Association. J. F. Oberlin University. Sep 3-4.

Yamada, A. 2022b. Tekiyokei no Tuuziteki Koubun Koutai: "Saseteitadaku," "Sasetemorau." "Sasetekudasaru" and "Sasetekureru" no Sentaku nitaisuru Zyootai Kuukan Moderu o Motiita Zikeiretu Bunseki [Diachronic Constructional Alternation of High-Applicative Forms: A Time-Series Analysis Using a State-Space Model for the Choice among "*Sasete itadak*," "*Sasete moraw*," "*Sasete kudasar*," and "*Sasete kure*"]. Oral Presentation at the 66rd Meeting of the Mathematical Linguistic Society of Japan. Sep 17, Tokyo, Nihon University.

編集後記

『日本語用論学会　第25 回大会発表論文集』第18号をお届けいたします。日本語用論学会では、2005年度より年次大会でのご発表内容を論文集としてとりまとめ、大会後に発行しております。今号では、シンポジウム3件、研究発表27件（日本語発表23件、英語発表7件）、合計30件のご寄稿をいただきました。 第26回大会後は『日本語用論学会　第26回大会発表論文集』第19号を発行する予定でございますので、どうぞご期待ください。

＊従来、巻末に掲載しておりました 日本語用論学会規約 は、紙面削減のため、割愛させていただきました。学会サイト（http://www.pragmatics.gr.jp）をご覧ください。

（『大会発表論文集』編集担当　大志民彩加・中馬隼人・八木橋宏勇）

日本語用論学会　第25回大会発表論文集　第18号（2022）
(Proceedings of the 25th Conference of the Pragmatics Society of Japan)

発　行　日　2023年8月31日
代　表　者　滝浦　真人
編集・発行　日本語用論学会　事務局（The Pragmatics Society of Japan）
〒560-0043 大阪府豊中市待兼山町1-8
大阪大学大学院人文学研究科言語文化学専攻
秦かおり　研究室内
E-mail:　secretary@pragmatics.gr.jp
印　　　刷　（株）田中プリント
〒600-8047　京都市下京区松原通麩屋町東入ル677-2
TEL：075-343-0006　FAX：075-341-4476