「研究ノート」

Measuring the Effects of Frequency and Dispersion on Key Expression Analysis: Methodological Recommendations

Yuichiro KOBAYASHI

Abstract

The automatic identification of key expressions is among the most crucial tasks in corpus linguistics. The use of random forest (RF) for analyzing key expressions has been increasing lately owing to its ability to efficiently compute the keyness scores of many linguistic features. However, only little is known of the types of predictor variables that RF assigns high scores to. It is significantly risky to rely on a single method without understanding its pros and cons. Thus, this study was conducted to distinguish between the two most employed machine-learning algorithms for variable selection – RF and least absolute shrinkage and selection operator (LASSO) – as the tools for extracting key expressions from corpora. Specifically, employing the corpus of *PLOS ONE* research articles, both statistical methods were compared regarding the effects of frequency and dispersion. The results indicate that RF selects the high-frequency variables in a large number of texts, while LASSO selects the low-frequency variables in a small number of texts. Since both methods have their pros and cons, the purpose of the research will determine the method to be adopted.

1. Introduction

Corpus analysis generally involves the comparison of multiple text groups, as well as the extraction of key expressions, such as keywords and key *n*-grams, that characterize each group (Scott & Tribble, 2006). Specifically, corpus-based stylistic studies differentiate the language use of a particular writer or fictional character from the linguistic *norms* represented by general or reference corpora (Stubbs, 2005). By comparing learners at different proficiency levels, learner corpus studies also reveal

their characteristics at each level (Pérez-Paredes & Díez-Bedmar, 2019). Furthermore, studies on English for specific purposes have produced a list of the key expressions in a particular academic field by analyzing the corpus of research articles on multiple academic disciplines (Asano, 2018). Regardless of the differences in the analyzed registers, these studies were all aimed at identifying the key expressions that are more salient in a group of texts than in others.

Various statistical methods have been employed to identify such key expressions. Conventionally, most analyses of key expressions have relied on the log-likelihood ratio and chi-square tests that are implemented in most corpus analysis tools. However, these tests cannot account for the variance within a group since the individual texts within a group are aggregated at the group level. For example, when distinguishing between the styles of two writers, the characteristics of their individual texts are generally ignored, and the writers are directly compared. This represents a limitation that several researchers have attempted to overcome via mean and median tests (Paquot & Bestgen, 2009) or dispersion measures (Egbert & Biber, 2019; Gries, 2021) for keyword extraction.

A recent trend in keyness analysis is the application of machine-learning methods in the detection of linguistic features that can predict text groups or linguistic choices. Particularly, the utilization of random forest (RF) models (Breiman, 2001) has been increasing in several areas of corpus linguistics, such as grammatical studies (Deshors, 2019; Deshors & Gries, 2016; Hundt et al., 2020; Paquot et al., 2019), sociopragmatics (Funke & Bernaisch, 2022), stylometry (Suzuki & Hosoya, 2014; Tabata, 2014), and learner corpus research (Kobayashi & Abe, 2016; Kobayashi et al., 2022; Tono, 2013). RF can efficiently analyze thousands of linguistic features and compute variable importance scores for each feature, representing the magnitude of the differences between the frequencies of the compared text groups or linguistic choices.

However, it is dangerous to depend on a single method without understanding its advantages and disadvantages. RF models are not always stable or robust, and even small changes in the dataset can significantly change the results of the analysis (Gries, 2020). In addition, the changes in the method can also significantly impact the results of the data analysis (Silberzahn et al., 2018). In the field of corpus linguistics, it is well-known that the choice of collocation measures greatly affects the results of the collocation analysis. Generally, collocations with high *t*-scores exhibit high-frequency

pairs, whereas those with high mutual-information scores include low-frequency words (McEnery et al., 2006). This knowledge of collocation statistics reveals the risks of relying on a single method. Thus, it is very crucial to understand the differences in key expression analyses that are performed with different statistical methods.

In machine learning, key expression analysis is considered an application of variable selections that identifies crucial predictor variables for high-accuracy data classification. The two most widely employed algorithms for variable selection are RF and the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996). Both statistical methods have achieved high success in data analysis competitions, such as Kaggle, because they generally offer good predictive performance and low overfitting (Banerjee, n.d.). Owing to their abilities to analyze many variables efficiently, both methods also exhibit great potential in text analysis for measuring the keyness of all words or *n*-grams in the corpus.

2. RF and LASSO

RF consists of hundreds or thousands of decision trees; each tree is built via the random samplings of the observations from the dataset and predictor variables. The variables are not completely considered by every tree, which renders the trees decorrelated and less prone to overfitting. The RF model performs a final prediction by synthesizing the results of each tree using the majority vote. In this model, the mean decrease in the Gini coefficient is employed as a measure of the contribution of each variable to the prediction: the higher the value of the mean decrease in the Gini coefficient, the higher the significance (keyness) of the variable in the model.

LASSO is a modeling method that simultaneously performs variable selection and model building. Technically, this method employs L1 regularization, which adds a penalty that is equivalent to the absolute value of the magnitude of the coefficients (Hastie et al., 2009). By adding the penalty, LASSO can avoid overfitting and remove the insignificant predictor variables from the dataset. The resulting reduced number of variables enhances the prediction accuracy and interpretability of the model. In the LASSO model, the partial regression coefficients of each variable can be used as the keyness scores for the key expression analysis.

3. Purpose of the Study

This study was conducted to compare two statistical methods, RF and LASSO, as tools for extracting key expressions from corpora. Particularly, both methods were compared in terms of frequency and dispersion. In this study, the term *key expressions* is defined as linguistic features that are instrumental in the statistical discrimination of two or more text groups. The term *frequency* indicates the number of times an expression is used in the text, and *dispersion* refers to the number of texts a given expression appeared in. The following research question was explored here: How do RF and LASSO differ regarding the effects of frequency and dispersion? The answer to this question can offer methodological recommendations with which corpus linguists can select appropriate keyness measures for their research.

4. Corpus

In this study, a corpus of 1,000 *PLOS ONE* research articles was employed on machine learning. The corpus was compiled with AntCorGen (version 1.2.0) (Anthony, 2022), which was the latest version at the time of conducting the study. To extract the key expressions from the corpus, the introduction (INT) and results and discussion (RAD) sections of the 1,000 articles were compared. Table 1 reveals the numbers of texts and words in each section analyzed in this study.

Table 1. Numbers of texts and words per section

	Number of texts	Number of words
Introduction (INT)	1,000	1,403,982
Results and discussion (RAD)	1,000	4,486,595
Total	2,000	5,890,577

5. Data Analysis

In this study, the relative frequencies (per 100 words) of 65 of 67 Biber's (1988) linguistic features were counted using the Multidimensional Analysis Tagger (Nini, 2019). The values of the remaining two variables – type/token ratio (TTR) and mean word length (AWL) – were also calculated by the tool. Thereafter, these 67 values were

employed to compare the INT and RAD sections. All statistical analyses were conducted using R (version 4.2.0), a free software environment for statistical computation and graphics (R Core Team, 2021). The *randomForest* and *glmnet* packages were used to perform the RF and LASSO analyses, respectively. Further, rank sum tests, correlation analysis, and linear regression analysis were conducted to complement the results of RF and LASSO.

6. Results and Discussion

6.1 Key Expression Analysis via RF

This study began with the key expression analysis using RF. The RF model employed Biber's 67 linguistic features as the predictor variables and utilized two text types (INT and RAD) as the response variables; the results of the classification are presented in Table 2, in which the columns and rows of the matrix represent the text types predicted by the model and the actual text types, respectively. Thus, the column "Accuracy" indicates the agreement rates between the predicted and actual text types. In this classification model, high accuracy indicates significant differences in the frequency patterns of the linguistic features between the INT and RAD sections.

Table 2. Accuracy rates of the RF models of the two text types

	INT (predicted)	RAD (predicted)	Accuracy
INT (actual)	956	44	95.6%
RAD (actual)	26	974	97.4%

Note. Overall accuracy rate evaluated via out-of-bag simulation was 96.5%.

RF can be used to compute the keyness scores, also known as the variable importance scores, for measuring the impact of each predictor variable on the alternation, given all the other predictors. The keyness scores demonstrated that the top-five linguistic features that distinguished the text types were (a) PEAS (perfect aspect), (b) TO (infinitives), (c) VPRT (present tense), (d) AWL (mean word length), and (e) VBD (past tense). The Wilcoxon rank sum tests with continuity correction also indicated that significant differences existed between the INT and RAD sections of all five variables ($p < 0.001$). One of the easiest strategies for identifying the linguistic

features that characterize each section is to draw their box plots. Figure 1 shows the box plots of the frequency distributions of the top-five linguistic features. The plots revealed that PEAS, TO, VPRT, and AWL were characteristic of INT, while VBD was characteristic of RAD.

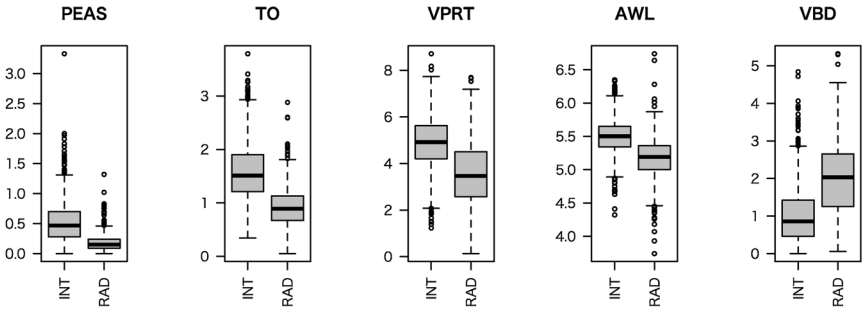

Figure 1. Frequency distributions of the top-five linguistic features in the RF model

The key expression analysis using RF is generally performed via the aforementioned procedure. However, the types of predictor variables that RF assigns high scores to are barely known. Therefore, this study was aimed at elucidating the relationships among the keyness scores, frequencies (or the magnitude of the values), and dispersion values of each predictor variable. Figure 2 visualizes the relationship between the keyness scores and relative frequency, as well as between the keyness scores and dispersion in the RF model. The horizontal and vertical axes of the figures indicate the logarithm of the keyness scores and frequency and dispersion values of the 67 linguistic features, respectively. The straight lines, representing the results of the linear regression analysis, indicated that RF tends to assign high scores to frequently used very dispersed variables. Pearson's product-moment correlation coefficients also revealed high, positive correlations between the keyness scores and frequency values (r = 0.57) and between the keyness scores and dispersion values (r = 0.65). Put differently, RF tends not to extract key expressions that are prominently utilized in a particular text type, albeit at low frequency or only in a few texts. This tendency is generally observed in other decision-tree-based models including gradient boosting (Friedman, 2001). Moreover, RF generates keyness scores for all predictor variables, although no theoretical threshold can be used to discriminate between the relevant and

irrelevant variables. Thus, the threshold setting generally tends to become arbitrary. Despite these shortcomings, this method can be useful for studies of academic English that identify frequent words and phrases in each section of academic articles. The method also can be instrumental in contrasting the language use of learners at different proficiency levels.

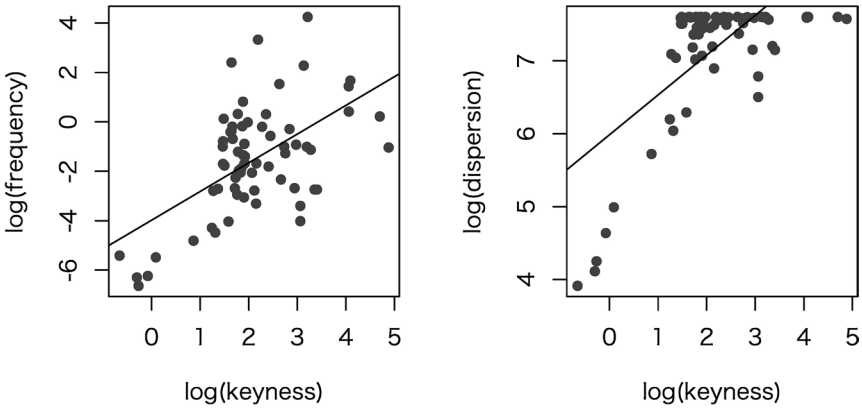

Figure 2. Relationships among the keyness, relative frequency, and dispersion in the RF model

6.2 Key Expression Analysis via LASSO

Owing to the threshold-setting issue in RF, LASSO was considered as the tool for key expression analysis, and Table 3 presents the classification results of the LASSO model using 67 linguistic features and two text types as the predictors and responses, respectively. Since the overall accuracy rate was 95.6%, modeling with LASSO was as highly reliable as that with RF (96.5%).

Employing the variable selection process in the LASSO model, 25 of the 67 predictor variables were statistically selected. The keyness scores (i.e., the absolute

Table 3. Accuracy rates of the LASSO model for the two text types

	INT (predicted)	RAD (predicted)	Accuracy
INT (actual)	952	48	95.2%
RAD (actual)	40	960	96.0%

Note. Overall accuracy rate evaluated via cross validation was 95.6%.

values of the coefficients) indicated that the top-five linguistic features that can distinguish text types were (a) SMP (*seem* and *appear*), (b) THAC (*that* adjective complements), (c) DPAR (discourse particles), (d) TOBJ (*that* relative clauses on object position), and (e) STPR (stranded prepositions). These five linguistic features do not overlap with a single feature in the top-five features of the RF model. Figure 3 shows the frequency distributions of the top-five variables in the LASSO model. The values on the vertical axes indicate that the low-frequency variables were assigned high scores. Additionally, the medians of one or both text types for the five variables were zero. However, the Wilcoxon rank sum tests with continuity correction detected significant differences between the text types for all five variables ($p < 0.001$).

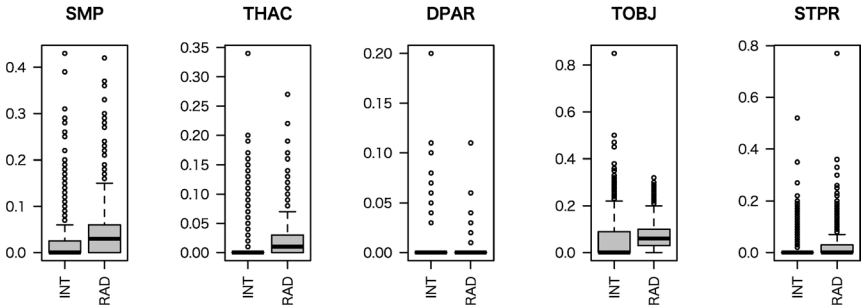

Figure 3. Frequency distributions of the top-five linguistic features in the LASSO model

Figure 4 shows the relationship between the keyness and relative frequency and between the keyness and dispersion in the LASSO model. Pearson's product-moment correlation coefficients also indicated the high, negative correlations between the keyness scores and frequency values ($r = -0.66$) and between them and the dispersion values ($r = -0.65$). The results indicate that, compared with RF, LASSO tends to assign high scores to variables that are used at low frequencies in a small number of texts. From a linguistics standpoint, while the RF model highlighted the difference in frequency of features such as tense and aspect markers (PEAS, VPRT and VBD), the LASSO model emphasized the difference in features such as stance markers (SMP) and discourse markers (DPAR). Owing to its nature, in stylometry, LASSO can identify the specific linguistic features that an author uses in comparison to other authors.

Furthermore, in the studies of English for specific purposes, this method can detect idiosyncratic words and phrases in a particular academic or professional field.

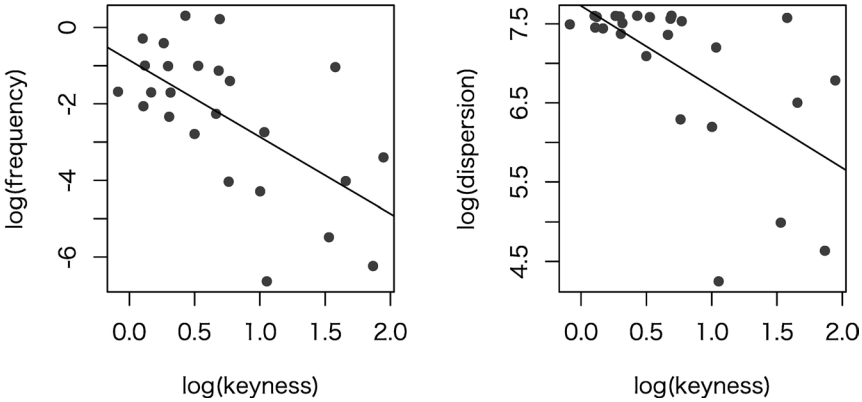

Figure 4. Relationships among the keyness scores, relative frequency, and dispersion in the LASSO model

7. Conclusion

This study was conducted to compare RF and LASSO as tools for analyzing key expressions from frequency and dispersion viewpoints. The results indicate that the keyness scores of RF and LASSO are positively and negatively related to the frequency and dispersion values, respectively. In many cases, RF models with a tendency to select high-frequency variables may be easier to interpret than LASSO models. However, the high-frequency features might be predictable without performing a statistical analysis. Additionally, as previously mentioned, the models cannot explicitly distinguish relevant variables from irrelevant ones. Conversely, the LASSO models with a tendency to select low-frequency variables could offer new insights that transcend the analyst's predictions. Furthermore, the LASSO models can remove irrelevant variables from the analysis. Nevertheless, the models only select one of the variables when there is a pair of highly correlated variables (Freijeiro-González et al., 2022). Therefore, if no substantial difference exists between the classification accuracies of both algorithms, the choice between them depends on the purposes of the intended studies. To deeply understand the differences between RF and LASSO, the methods must be compared

via various types of datasets and measures other than frequency and dispersion. Additionally, it is interesting to consider several improved versions of the algorithms, such as Boruta (Kursa & Rudnicki, 2010) and elastic net (Zou & Hastie, 2005), as tools for comparison. Boruta can clearly distinguish between relevant and irrelevant variables by embedding statistical tests in the RF model. Elastic net can assign the same keyness scores to highly correlated variables by combining LASSO and ridge regression models. Moreover, other statistical techniques than the RF- and LASSO-related methods should be applied to seek a better set of key expressions. Similar to the findings on the collocation statistics, the knowledge of the appropriate use of multiple keyness measures can increase the validity of statistical analysis in corpus linguistics.

Acknowledgements

This research has been partly funded by Grants-in-Aid for Scientific Research (Grant Number: 21K00660). I would like to express my gratitude to the two anonymous reviewers for their helpful comments and suggestions.

References

Anthony, L. (2022). AntCorGen (Version 1.2.0). Waseda University. https://www.laurenceanthony.net/software

Asano, M. (2018). Construction of medical research article corpora with AntCorGen: Pedagogical implications. *English Corpus Studies, 25,* 101–115.

Banerjee, P. (n.d.). Comprehensive guide on feature selection. Kaggle. https://www.kaggle.com/code/prashant111/comprehensive-guide-on-feature-selection

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Deshors, S. C. (2019). English as a lingua franca: A random forests approach to particle placement in multi-speaker interactions. *International Journal of Applied Linguistics, 30*(2), 214–231.

Deshors, S. C., & Gries, S. Th. (2016). Profiling verb complementation constructions across New Englishes: A two-step random forests analysis to *ing* vs. *to* complements. *International Journal of Corpus Linguistics, 21*(2), 192–218.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104.

Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review, 90*(1), 118–145.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Funke, N., & Bernaisch, T. (2022). Intensifying and downtoning in South Asian Englishes: Empirical perspectives. *English World-Wide, 43*(1), 33–65.

Gries, S. Th. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory, 16*(3), 617–647.

Gries, S. Th. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1–33.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hundt, M., Rautionaho, P., & Strobl, C. (2020). Progressive or simple? A corpus-based study of aspect in World Englishes. *Corpora, 15*(1), 77–106.

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics, 20*(1), 55–73.

Kobayashi, Y., Abe, M., & Kondo, Y. (2022). Exploring L2 spoken developmental measures: Which linguistic features can predict the number of words? *English Corpus Studies, 29,* 1–18.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13.

McEnery, A., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Routledge.

Nini, A. (2019). The multi-dimensional analysis tagger. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury Academic.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 247–269). Rodopi.

Paquot, M., Grafmiller, J., & Szmrecsanyi, B. (2019). Particle placement alternation in EFL learner vs. L1 speech: Assessing the similarity of probabilistic grammars. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research: Selected papers from the Fourth Learner Corpus Research Conference* (pp. 71–92). Presses universitaires de Louvain.

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2019). Researching learner language through POS keyword and syntactic complexity analyses. In S. Götz & J. Mukherjee (Eds.), *Learner corpora and language teaching* (pp. 101–128). John Benjamins.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.r-project.org/

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language*

education. John Benjamins.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1,* 337–356.

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature, 14*(1), 5–24.

Suzuki, T., & Hosoya, M. (2014). Computational stylistic analysis of popular songs of Japanese female singer-songwriters. *Digital Humanities Quarterly, 8*(1). http://www.digitalhumanities.org/dhq/vol/8/1/000170/000170.html

Tabata, T. (2014). Stylometry of Dickens's language: An experiment with random forests. In P. L. Arthur & K. Bode (Eds.), *Advancing digital humanities: Research, methods, theories* (pp. 28–53). Palgrave Macmillan.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58,* 267–288.

Tono, Y. (2013). Criterial feature extraction using parallel learner corpora and machine learning. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 169–204). John Benjamins.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, 67*(2), 301–320.

Appendix: The 67 linguistic features from Biber (1988)

A. Tense and aspect markers

1. past tense (VBD), 2. perfect aspect (PEAS), 3. present tense (VPRT)

B. Place and time adverbials

4. place adverbials (PLACE), 5. time adverbials (TIME)

C. Pronouns and pro-verbs

6. first person pronouns (FPP1), 7. second person pronouns (SPP2), 8. third person personal pronouns (excluding *it*) (TPP3), 9. pronoun *it* (PIT), 10. demonstrative pronouns (DEMP), 11. indefinite pronouns (INPR), 12. pro-verb *do* (PROD)

D. Questions

13. direct WH-questions (WHQU)

E. Nominal forms

14. nominalizations (ending in *-tion, -ment, -ness, -ity)* (NOMZ), 15. gerunds (GER), 16. total other nouns (NN)

F. Passives

17. agentless passives (PASS), 18. *by*-passives (BYPA)

G. Stative forms

19. *be* as main verb (BEMA), 20. existential *there* (EX)

H. Subordination features

21. *that* verb complements (THVC), 22. *that* adjective complements (THAC), 23. WH clauses (WHCL), 24. infinitives (*to*-clause) (TO), 25. present participial clauses (PRESP), 26. past participial clauses (PASTP), 27. past participial WHIZ deletion relatives (WZPAST), 28. present participial WHIZ deletion relatives (WZPRES), 29. *that* relative clauses on subject position (TSUB), 30. *that* relative clauses on object position (TOBJ), 31. WH relatives on subject position (WHSUB), 32. WH relatives on object position (WHOBJ), 33. pied-piping relative clauses (PIRE), 34. sentence relatives (SERE), 35. causative adverbial subordinators (*because*) (CAUS), 36. concessive adverbial subordinators (*although, though*) (CONC), 37. conditional adverbial subordinators (*if, unless*) (COND), 38. other adverbial subordinators (OSUB)

I. Prepositional phrases, adjectives, and adverbs

39. total prepositional phrases (PIN), 40. attributive adjectives (JJ), 41. predicative adjectives (PRED), 42. total adverbs (RB)

J. Lexical specificity

43. type/token ratio (TTR), 44. mean word length (AWL)

K. Lexical classes

45. conjuncts (CONJ), 46. downtoners (DWNT), 47. hedges (HDG), 48. amplifiers (AMP), 49. emphatics (EMPH), 50. discourse particles (DPAR), 51. demonstratives (DEMO)

L. Modals

52. possibility modals (POMD), 53. necessity modals (NEMD), 54. predictive modals (PRMD)

M. Specialized verb classes

55. public verbs (PUBV), 56. private verbs (PRIV), 57. suasive verbs (SUAV), 58. *seem* and *appear* (SMP)

N. Reduced forms and dispreferred structures

59. contractions (CONT), 60. subordinator *that* deletion (THATD), 61. stranded prepositions (STPR), 62. split infinitives (SPIN), 63. split auxiliaries (SPAU)

O. Coordination

64. phrasal coordination (PHC), 65. independent clause coordination (ANDC)

P. Negation

66. syntactic negation (SYNE), 67. analytic negation (XX0)

（小林雄一郎　日本大学）