A Machine Learning Approach to the Effects of Writing Task Prompts

Kobayashi, Yuichiro

Abe, Mariko

(Citation)
Learner Corpus Studies in Asia and the World,2:163-175

(Issue Date)
2014-05-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81006698

(URL)
https://hdl.handle.net/20.500.14094/81006698

A Machine Learning Approach to the Effects of Writing Task Prompts

Yuichiro KOBAYASHI
Japan Society for the Promotion of Science

Mariko ABE
Chuo University

Abstract

In analyzing learner language use, it is important to understand the influence of prompts given for writing tasks. The study reported upon in the present paper aimed to investigate the effects of essay prompts on language use in essays written by English as a foreign language students, using a machine learning approach. The study compared the essays of East-Asian learners of English in terms of various linguistic features (e.g., parts-of-speech, grammar, and discourse features) proposed by Biber (1988). The essays were drawn from the International Corpus Network of Asian Learners of English (Ishikawa, 2011), which is considered to be the largest East Asian database of written compositions. The database contains essays written in response to two prompts, namely (a) *It is important for college students to have a part time job*, and (b) *Smoking should be completely banned at all the restaurants in the country*. The results of the present study indicate that the language used in the prompts significantly influenced the learners' language production. The findings suggest that it is crucial to consider essay prompts in terms of parts-of-speech, grammar, and discourse features, and to investigate the influence of such essay prompts on written compositions when conducting learner corpus studies.

Keywords

L2 writing, Wording of writing task prompts, Machine learning

I Introduction

A learner corpus is a computerized textual database of language produced by foreign language learners (Leech, 1998). Construction of such learner corpora has flourished in the last two decades (e.g., Pravec, 2002), allowing a new trend in research on such learners wherein multiple learner corpora are compared (e.g., Granger, 2013; Tono,

2007). In general, a corpus is constructed with a specific purpose on the basis of explicit design criteria (Atkins, Clear, & Ostler, 1992), and the results of analyses in corpus linguistics depend on the corpus itself (e.g., Sinclair, 1991). Therefore, it is crucial for researchers to consider the design of a particular corpus when making comparisons across different corpora. In particular, topic, task, and text length can have a possibility to influence the "opportunity of use" for a given linguistic feature (Buttery & Caines, 2012).

II Research Design

2.1 The Purpose of the Study

The present study aimed to investigate the effects of essay prompts on language use in the essays of English as a foreign language (EFL) learners, using a machine learning approach. The study compares language use in the essays in terms of a number of linguistic features originally examined by Biber (1988), namely parts-of-speech, grammar, and discourse features. It also attempts to identify particular linguistic features whose frequency of use may be affected by a difference in the prescribed essay topic.

2.2 Corpus Data

This study draws on the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011), a corpus of one million words in written samples from EFL learners and native speakers, which is considered to be the largest East Asian composition database. The corpus contains essays written in response to two different prompts, namely (a) *It is important for college students to have a part time job (PTJ)*, and (b) *Smoking should be completely banned at all the restaurants in the country (SMK)*. Students are required to write their essays using word-processing software and without the use of dictionaries and other reference tools. They are asked to do an electronic spell-check before submitting the essay. The essays are required as extra homework for their EFL classes. The data analyzed in the present study is a subset from this corpus, including the written compositions of 2,000 EFL learners. Table 1 shows the size of each sub-corpus, according to essay topic and the learners' country of origin. The number of students from each country who wrote on each topic is identical. The length of each essay was between 200 and 300 words.

Table 1 Corpus size of four groups of East Asian EFL learners

	Hong Kong (HK)	Korea (KOR)	Taiwan (TWN)	Japan (JPN)	Total
Participants (PTJ)	100	300	200	400	1,000
Participants (SMK)	100	300	200	400	1,000
Total words	47,365	135,447	91,497	177,236	451,545
Mean words per essay	237	226	229	222	226

2.3 Linguistic Features

As pointed out in Biber (1986) and Biber, Conrad, and Reppen (1998), limiting the number of linguistic features targeted in a corpus study can yield limited results. The 67 linguistic features proposed by Biber (1988) are broadly used in the field of corpus linguistics to examine variation in linguistic texts (e.g., Conrad & Biber, 2001; Frignal, 2013). Applying these features to learner corpora in previous research, we have succeeded in identifying the linguistic features that can be used to discriminate between different East Asian EFL learners and to distinguish native and non-native speakers of English (Abe, Kobayashi, & Narita, 2013). The present study applied 58 of the 67 linguistic features[1] examined by Biber (1988) to the EFL essay corpus. A full list of the linguistic features analyzed in the present study can be found in the Appendix.

2.4 Statistical Methods

We employed an algorithm of machine learning, namely random forests (Breiman, 2001; Breiman & Culter, n.d.). Random forests can be defined as an ensemble learning method that operates by constructing a large collection of decision trees. The method generally achieves higher levels of accuracy than other statistical classifiers. Moreover, it can handle thousands of input variables without leading to variable deletion, thereby providing reliable estimates of what variables are important for the classification.

2.5 Procedures

We began by calculating the frequencies of the 58 linguistic features in the 2,000 essays. Using random forests, we then compared the 1,000 PTJ essays with the 1,000 SMK essays. Following this comparison, we identified linguistic features that were characteristic of essays written on each topic. Finally, we examined the similarity of the frequency patterns of linguistic features in the essay prompts and in the learners' essays.

III Results and Discussion

We began by examining whether language use, as reflected by the 58 linguistic features, differed between the PTJ and SMK essay topics. Random forests allow a visual

representation of the similarity between texts in a multidimensional scaling diagram. The similarities across the 2,000 essay writings are shown in Figure 1. Additionally, the error rates for predicting the essay topic from the pattern of linguistic features using the statistical model are provided in Table 2.

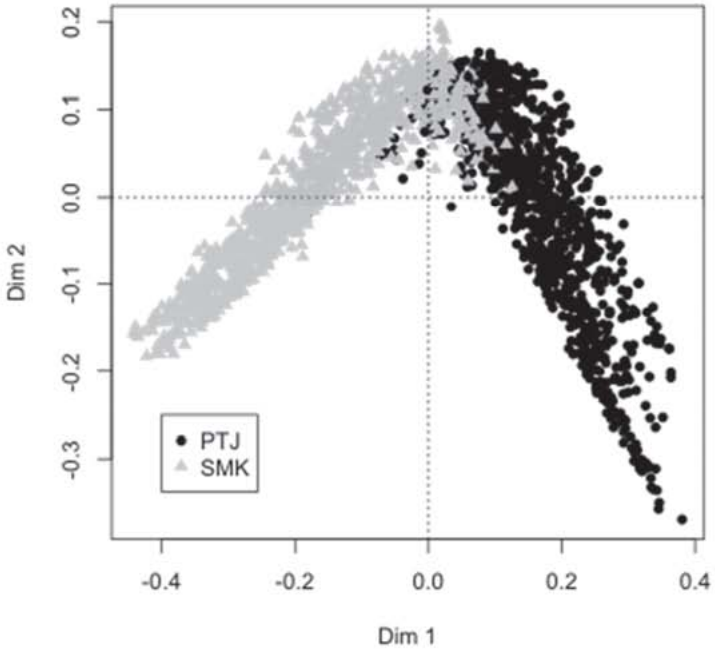

Fig. 1 Multidimensional scaling diagram showing the similarities among 2,000 essays

Table 2 Error rates of random forests for PTJ and SMK

	PTJ	SMK	Error rate
PTJ	943	57	0.057
SMK	119	881	0.119

Note. Overall error rate was 0.088.

In Figure 1, each dot represents an essay on the PTJ topic and each triangle represents an essay on the SMK topic. The coordinates in the diagram reflect the relationships between essays on the two topics. The distribution of the essays shows that those on each topic cluster together, with PTJ essays clustered on the right-hand side of the diagram and SMK essays on the left. This clustering suggests that essays on the two topics differ in their use of the 58 linguistic features.

Table 2 shows how accurately the random forests model is able to classify the essays

in terms of topic. The model yielded the correct prediction for 943 of the 1,000 PTJ essays (reflecting an error rate of 5.7%), and for 881 of the 1,000 SMK essays (reflecting an error rate of 11.9%). The average error rate was 8.8%, showing that the random forests model could correctly predict the topic of 91.2% of the 2,000 essays.[2] These findings suggest a substantial difference in language use between PTJ and SMK essays.

The next step was to identify linguistic features characteristic of each essay topic. One of the merits of random forests in this study was its ability to highlight the effect of the essay prompt on the occurrence of the linguistic features by means of the *MeanDecreaseGini*. This measure indicates the mean decrease in the Gini index, which is an index of the uneven distribution of a particular variable between groups when the variable in question is excluded from analysis (Tabata, 2012). Figure 2 presents the *MeanDecreaseGini* for the 58 linguistic features.

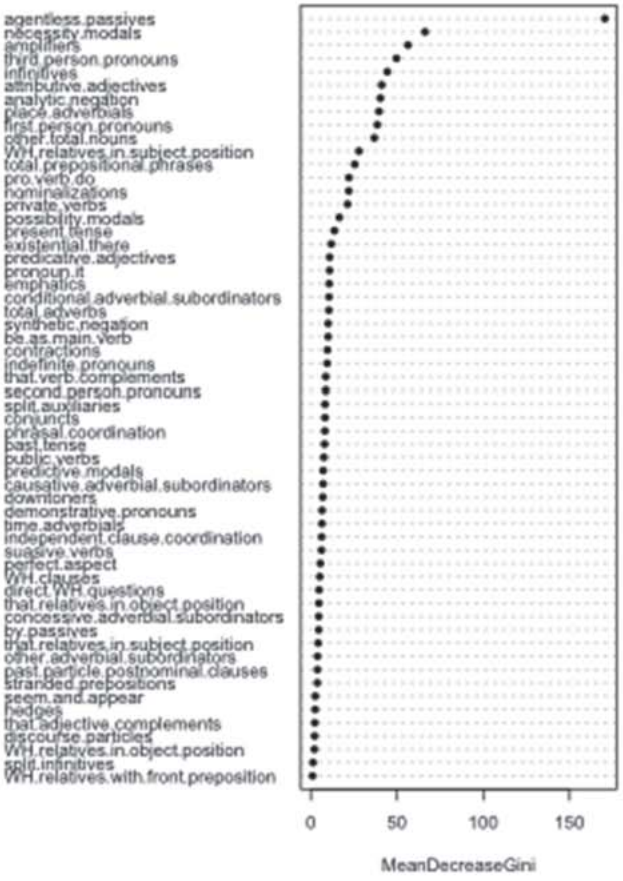

Fig. 2 Linguistic features sorted by *MeanDecreaseGini*

As is clear from Figure 2, the top 10 linguistic features that are able to discriminate essay topics were (a) agentless passives, (b) necessity modals, (c) amplifiers, (d) third person pronouns, (e) infinitives, (f) attributive adjectives, (g) analytic negation, (h) place adverbials, (i) first person pronouns, and (j) other total nouns. Whereas the *MeanDecreaseGini* identifies the linguistic features that are heavily affected by the essay prompt, it cannot specify which essay prompt had a strong association with a given linguistic feature. It is therefore informative to consider a graphic depiction of the marginal effect of a variable on the class probability. Figure 3 shows the relationship of each of the top 10 linguistic features and essay topics in partial dependency plots.

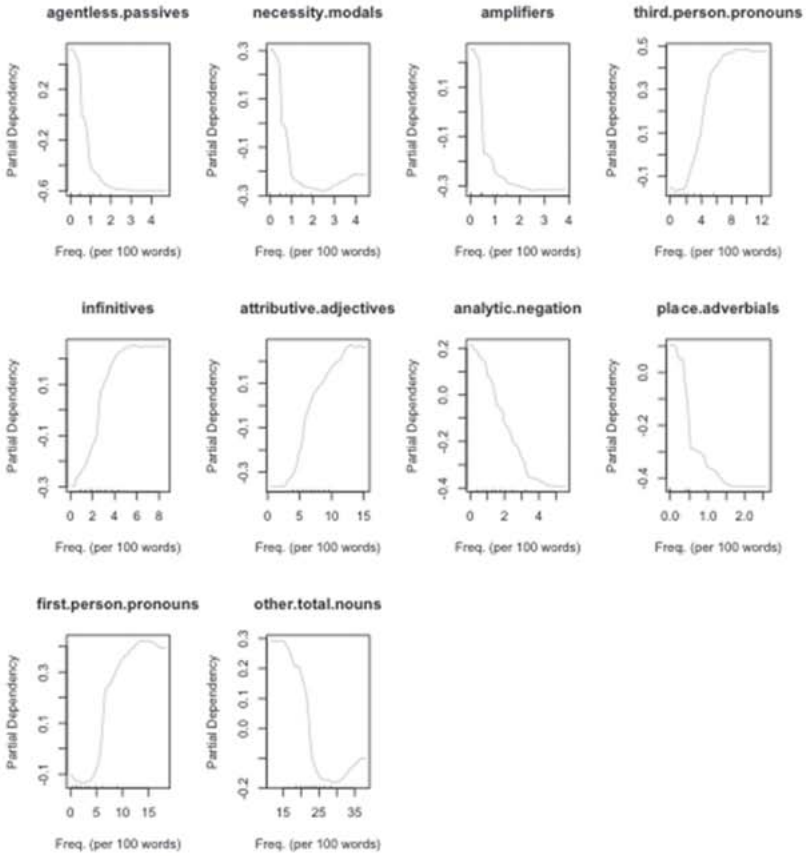

Fig. 3 The relationships between the top 10 linguistic features and essay topics

The horizontal axes in these partial dependency plots indicate the frequency of the

particular linguistic feature (per 100 words). The vertical axes show the probability of essay topic. A rising line indicates that the particular linguistic feature was frequently used in PTJ essays, and a falling line indicates frequent use in SMK essays. Thus, these plots show which linguistic features were more commonly used in essays on each topic. In addition, the frequency distributions of the top 10 linguistic features are shown in the form of box plots in Figure 4.

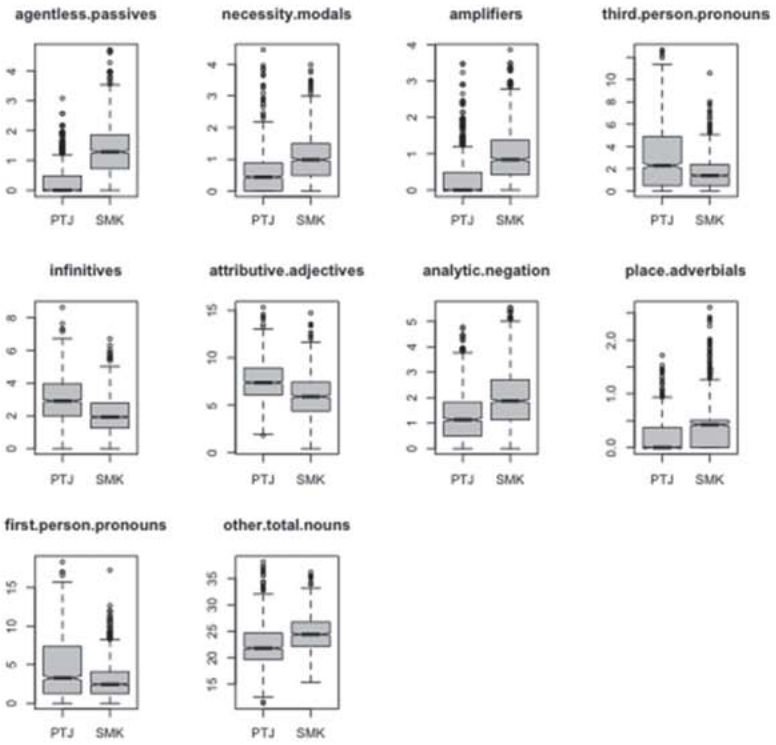

Fig. 4 Frequency distributions of the top 10 linguistic features (per 100 words)

According to the box plots, four linguistic features were frequently used in PTJ essays, as compared to SMK essays, namely (a) third person pronouns, (b) infinitives, (c) attributive adjectives, and (d) first person pronouns. Among these four features, both infinitives and attributive adjectives were used in the prompt for PTJ essays (*It is important for college students to have a part time job*). On the other hand, the six remaining linguistic features were more frequently used in SMK essays, namely (a) agentless passives, (b) necessity modals, (c) amplifiers, (d) analytic negation, (e) place

adverbials, and (f) other total nouns. Among these six features, agentless passives, necessity modals, amplifiers, and other total nouns were used in the prompt for SMK essays (*Smoking should be completely banned at all the restaurants in the country*). Although other total nouns were used in both essay prompts, they were frequently used in SMK essays. Thus, the analysis reveals that some of the linguistic features that were most frequently used in essays by language learners were identical to those used in the essay prompt. This suggests that the reuse of the wording of an essay prompt may affect the language use in written compositions, and this result supports the n-gram analysis of Ishikawa (2009).

A further important finding is that agentless passives were the most discriminant linguistic feature in distinguishing the topic of essays in this study (see Figure 2). Passives are often used in calculations of text readability (e.g., Stockmeyer, 2009), as well as for automated essay evaluation (e.g., Landauer, Laham, & Foltz, 2003).[3] Thus, the frequency of passive constructions can potentially affect the score assigned to an essay. The use of passives in the wording of an essay prompt ought therefore to be carefully considered. In addition, the possible effects of other linguistic features in essay prompts, such as parts-of-speech (e.g., nouns), grammar (e.g., passives), and discourse features (e.g., amplifiers), should be carefully considered in terms of their possible effect on the written compositions of EFL learners.

IV Conclusion

The purpose of the present study was to investigate the effects of essay prompts on language use in EFL compositions using a machine learning approach. The findings suggest that the essays of EFL learners are greatly influenced by the nature of the essay prompt, and that essays written in response to different prompts differ in terms of the frequency of occurrence of certain linguistic features. Furthermore, the linguistic features occurring in the prompt were directly reflected in the learners' essays. These findings show that it is crucial to consider aspects of parts-of-speech, grammar, and discourse in essay prompts when conducting research involving learner corpora. Topic variety is strictly controlled in corpora like the ICNALE, but it is still necessary to pay attention to the language use in the wording of the essay prompt. In conclusion, it is important to remain aware of the potential influence of an essay prompt on learner essays in a corpus study.

Notes

[1] Seven features were excluded because of differences in the software used to annotate part-of-speech tags, namely (a) demonstratives, (b) gerunds, (c) present participial clauses, (d) past participial clauses, (e) present participial WHIZ deletion relatives, (f)

sentence relatives, and (g) subordinator-that deletion. In addition, two features, namely type/token ratio and word length, were excluded because they are categorized as word variation rather than as linguistic features.

[2] In random forests, there is no need for cross-validation or a separate test set to gain an unbiased estimate of the test set error. In the out-of-bag (OOB) error estimate, each tree is constructed using a different bootstrap sample from the original data. Approximately one-third of the cases are left out of the bootstrap sample and are not used in the construction of the classification model (Breiman & Cutler, n.d.).

[3] Frequencies of linguistic features, such as agentless passives and attributive adjectives, are associated with higher essay scores (Weigle, 2013).

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research Grant Numbers 12J02669, 24320101, 24520631, and 26770205.

References

Abe, M., Kobayashi, Y., & Narita, M. (2013). Using multivariate statistical techniques to analyze the writing of East Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world – Vol.1* (pp. 55-65). Kobe: School of Language and Communication, Kobe University.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62,* 384-414.

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

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-23.

Breiman, L., & Cutler, A. (n.d.). Random forests. Retrieved from http://www.stat.berkeley.edu/~breiman/RandomForests/

Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi & M. Minegishi (Eds.), *Developmental and crosslinguistic perspectives in learner corpus research* (pp. 187-204). Amsterdam: John Benjamins.

Conrad, S., & Biber, D. (Eds.) (2001). *Variation in English: Multi-dimensional studies.* Harlow: Longman.

Frignal, E. (2013). Twenty-five years of Biber's multi-dimensional analysis:

Introduction to the special issue and an interview with Douglas Biber. *Corpora, 8*(2), 137-152.

Granger, S. (2013). The passives in learner English: Corpus insights and implications for pedagogical grammar. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world – Vol.1* (pp. 5-15). Kobe: School of Language and Communication, Kobe University.

Ishikawa, S. (2009). Phraseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. In K. Yagi & T. Kanzaki (Eds.), *Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan* (pp. 87-98). Hyogo: Kwansei Gakuin University Press.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning, and research* (pp. 3-11). Glasgow: University of Strathclyde Press.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). New York: Routledge.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xxii). London: Longman.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal, 26*, 81-114.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Stockmeyer, N. O. (2009). Using Microsoft Word's readability program. *Michigan Bar Journal, 88*(1), 46-47.

Tabata, T. (2012). 'Key' words and stylistic 'signatures': Textometry and random forests. In T. Tabata (Ed.), *Mining textual patterns* (pp. 45-64). Tokyo: Institute of Statistical Mathematics.

Tono, Y. (2007). The roles of oral L2 learner corpora in language teaching: The case of the NICT JLE Corpus. In M. C. Compoy & M. J. Luzón (Eds.), *Spoken corpora in applied linguistics* (pp. 163-179). Bern: Peter Lang.

Weigle, S. C. (2013). English as second language writing and automated essay evaluation. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). New York: Routledge.

Appendix: *Linguistic Features Analyzed in the Present Study*

Linguistic category and examples
A. Tense and aspect markers
1. past tense
2. perfect aspect
3. present tense
B. Place and time adverbials
4. place adverbials (e.g., *across, behind, inside*)
5. time adverbials (e.g., *early, recently, soon*)
C. Pronouns and pro-verbs
6. first person pronouns
7. second person pronouns
8. third person pronouns (excluding *it*)
9. pronoun *it*
10. demonstrative pronouns (*that, this, these, those*)
11. indefinite pronouns (e.g., *anybody, nothing, someone*)
12. pro-verb *do* (e.g., *the cat did it*)
D. Questions
13. direct WH-questions
E. Nominal forms
14. nominalizations (ending in *-tion, -ment, -ness, -ity*)
15. other total nouns (except for nominalizations)
F. Passives
16. agentless passives
17. *by*-passives
G. Stative forms
18. *be* as main verb
19. existential *there* (e.g., *there are several explanations ...*)
H. Subordination
H1. Complementation
20. *that* verb complements (e.g., *I said that he went*)
21. *that* adjective complements (e.g., *I'm glad that you like it*)
22. WH-clauses (e.g., *I believed what he told me*)
23. infinitives (*to*-clause)
H2. Participial forms
24. past participial postnominal (reduced relative) clauses (e.g., *the solution produced by this process*)
H3. Relatives

25. *that* relatives in subject position (e.g., *the dog that bit me*)
26. *that* relatives in object position (e.g., *the dog that I saw*)
27. WH relatives in subject position (e.g., *the man who likes popcorn*)
28. WH relatives in object position (e.g., *the man who Sally likes*)
29. WH relatives with fronted preposition (e.g., *the manner in which he was told*)

H4. Adverbial clauses

30. causative adverbial subordinators: *because*
31. concessive adverbial subordinators: *although, though*
32. conditional adverbial subordinators: *if, unless*
33. other adverbial subordinators: (having multiple functions) (e.g., *since, while, whereas*)

I. Prepositional phrases, adjectives, and adverbs

34. total prepositional phrases
35. attributive adjectives (e.g., *the big horse*)
36. predicative adjectives (e.g., *the horse is big*)
37. total adverbs (except conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers)

J. Lexical classes

38. conjuncts (e.g., *consequently, furthermore, however*)
39. downtoners (e.g., *barely, nearly, slightly*)
40. hedges (e.g., *at about, something like, almost*)
41. amplifiers (e.g., *absolutely, extremely, perfectly*)
42. emphatics (e.g., *a lot, for sure, really*)
43. discourse particles (e.g., sentence initial *well, now, anyway*)

K. Modals

44. possibility modals (*can, may, might, could*)
45. necessity modals (*ought, should, must*)
46. predictive modals (*will, would, shall*)

L. Specialized verb classes

47. public verbs (e.g., *acknowledge, admit, agree*)
48. private verbs (e.g., *anticipate, assume, believe*)
49. suasive verbs (e.g., *agree, arrange, ask*)
50. *seem* and *appear*

M. Reduced forms and dispreferred structures

51. contractions
52. stranded prepositions (e.g., *the candidate that I was thinking of*)
53. split infinitives (e.g., *he wants to convincingly prove that ...*)
54. split auxiliaries (e.g., *they are objectively shown to ...*)

N. Coordination

55. phrasal coordination (e.g., *NOUN* and *NOUN, ADJ* and *ADJ*)

56. independent clause coordination (clause initial *and*) (e.g., *It was my birthday and I was excited.*)

O. Negation

57. synthetic negation (e.g., *no answer is good enough for Jones*)

58. analytic negation: *not* (e.g., *that's not likely*)