A Memory-Sensitive Classification Model of Errors in Early Second Language Learning

Brendan Tomoschuk and Jarrett T. Lovelett
Department of Psychology
University of California, San Diego
9500 Gilman Drive, La Jolla, CA, 92093-0109
{btomosch, jlovelet}@ucsd.edu

Abstract

In this paper, we explore a variety of linguistic and cognitive features to better understand second language acquisition in early users of the language learning app Duolingo. With these features, we trained a random forest classifier to predict errors in early learners of French, Spanish, and English. Of particular note was our finding that mean and variance in error for each user and token can be a memory efficient replacement for their respective dummy-encoded categorical variables. At test, these models improved over the baseline model with AUROC values of 0.803 for English, 0.823 for French, and 0.829 for Spanish.

1 Introduction

Learning a new language is often a challenging task for adults. However, there are many linguistic and cognitive factors that can facilitate (or impair) acquisition of a non-native language, ranging from properties of the languages a learner already knows, to the methods and nature of study. Much work has sought to manipulate these factors in order to both further our understanding of the cognitive systems in play and facilitate learning.

Here, we present a model that explores these factors to predict outcomes for three populations of language learners that use Duolingo, a language learning app that gamifies lessons for a wide variety of to-be-learned languages. We start by describing the various features we developed from the data before describing the random forest model used and the subsequent outcomes.

2 Related Work

Little work has been done building predictive models of adult language acquisition, but many studies have explored the linguistic factors that impact vocabulary learning in a non-native language. Semantic properties of nouns, for example, have been found to impact word learning. Cognates, or words that overlap in form and meaning in both languages (e.g. *lemon* in English and *limón* in Spanish), have been shown to be easier to learn (de Groot & Keijzer, 2000). The same study showed that words that are rated as more concrete (*hat* as opposed to *liberty*) are easier to learn. While perhaps more surprising than the cognate result, this effect is often explained by the fact that more concrete words create more perceptual connections to their conceptual referents (it is easier to imagine a physical hat than the abstract concept of liberty), and it is therefore easier to connect new words to concepts via those connections.

There are likewise many factors than can hinder word learning. For example, interlingual homographs, or words that share surface form but have different meanings (*pan* as something to fry on in English and bread in Spanish) are harder to process and may therefore also be harder to learn (Dijkstra, Timmermans & Schriefers, 2000).

Beyond the linguistic particulars of individual words, the temporal dynamics of learning can powerfully moderate memory. One of the most well established results in cognitive psychology is that two repetitions of a to-be-learned item are best separated by some temporal gap, if the goal is long-term retention (Ebbinghaus 1885/1964, Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Donovan & Radosevich, 1999; T. D. Lee & Genovese, 1988). That is, given a fixed amount of available time to learn something, a learner is better off distributing that time over multiple learning sessions than cramming it all into a single session. Further, the more time that is allowed to pass before a learner encounters a previously learned item again, the longer into the future the learner can

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 231–239
New Orleans, Louisiana, June 5, 2018.

expect to retain that item (or equivalently, the greater the probability of successful retrieval of that item at a particular future time; but see Cepeda et al. 2008).

Over a century of research has shown this spacing effect to be robust across the human lifespan (e.g. Vander Linde, Morrongiello, & Rovee-Collier, 1985; Ambridge, Theakston, Lieven, & Tomasello, 2006; Carpenter, 2009; Cepeda et al., 2008; Balota, Duchek, & Paullin, 1989), over many varieties of learning tasks (Cepeda et al., 2006; Donovan & Radosevich, 1999; T. D. Lee & Genovese, 1988), and perhaps most strikingly, for nearly every inter-repetition temporal gap that has been investigated, from seconds (Ebbinghaus, 1964), to a range of days (e.g. Cepeda et al., 2008), to years (Bahrick & Phelphs, 1987).

Moreover, the advantage of spacing seems to be enhanced when combined with active retrieval from long-term memory (as compared to passive restudy), making it particularly well-suited to a microtesting-based learning platform like Duolingo (Carpenter & DeLosh, 2006; Cull, 2000; Karpicke & Roediger, 2007; Landauer & Bjork, 1978; Rea & Modigliani, 1985). Crucially for our present purpose, a number of studies have examined the efficacy of spaced repetition specifically in second language learning, where it seems to be effective at least for vocabulary, and perhaps for grammar as well, although further research is needed (for a review, see Ullman & Lovelett, 2018).

3 Data

The data were collected in 2017 from Duolingo, as part of the NAACL HLT 2018 Shared Task on Second Language Acquisition Modeling (SLAM, Settles, Brust, Gustafson, Hagiwara & Madnani, 2018). The data consisted of exercise and phrase level information for three populations of language learners in their first 30 days of using the app: English-speaking learners of Spanish and French as well as Spanish-speaking learners of English.

The data were split into a training set, which consisted of each user's first 80% of sessions, a development set (for testing model generalization before the test phase) that contained the next 10% of each user's data, and a test set that contained the final 10% of exercises for each user. The training data set consisted of 1,882,701 exercises in total (38.9% from learners of Spanish, 43.8% from learners of English and 17.3% from learners of French), while the development data contained 255,383 exercises (45.3% from learners of Spanish, 37.6% learners of from English and 17.1% from learners of French), and the test set contained 249,484 exercises (45.9% from learners of Spanish, 37.4% from learners of English and 16.7% from learners of French).

4 Features

Our approach to modeling errors in second language acquisition was driven primarily by two distinct bodies of research: linguistic effects in second language acquisition, and drivers of robust memory in general. As such we discuss each set of features separately.

4.1 Linguistic features

In this section, we describe the semantic and morpho-syntactic features added to the model. Values for tokens that were not in databases listed below were set to the mean of the feature.

Word length. Orthographic and phonological length (*orthoLength* and *phonLength* respectively) are predictive of word difficulty, and longer written or spoken words generally leave more room for potential errors (Baddeley, Thomon & Buchanan, 1975). Phonological length was taken from the CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012).

Word neighbors. A greater number of orthographic and phonological neighbors (*orthoNei* and *phonNei*) for a given word in both the to-be-learned and known languages might cause interference leading to errors. These data were also taken from the CLEARPOND database.

Word Frequency. The log transformed frequency (*logWordFreq)* of the English, Spanish and French words to be learned were also included as predictors, as well as the average log frequency of the phonological (*logOrthoNeiFreq*) and orthographic neighbors (*logPhonNeiFreq*) in the to-be-learned as well as known language.

Edit Distance. Because cognate status impacts language learning, the Levenshtein distance between a given token and its translation to user language (English for Spanish and French learners, and Spanish for English learners) was calculated by feeding single word translations through the Google Translate API and calculating edit distances between the translations. Cognates like *lemon*

and *limón* should have a short edit distance, while words like *boy* and *niño* will have relatively longer distances.

Interlingual homographs. Additionally, the interlingual homograph status for each token (whether or not the token shares its surface form with a translation of a different token) were added for each language by using the Google Translate API.

Morphological Complexity. As a proxy for how morphologically complex any given word is, the number of morphological features present in the given morphology columns were summed and treated as a proxy for morphological complexity (*morphoComplexity)*.

Concreteness. Mean and standard deviations for concreteness ratings were taken from Brysbaert, Warriner and Kuperman (2014) and added to the model.

4.2 Memory features

Repetition & Experience. Each instance (i.e., each token in each exercise for each user) was labeled with (1) the number of times the current user had encountered that token, up to and including the current instance (*nthOccurrence)* and (2) the number of instances the user had seen in total, up to and including the current instance (*userTrial)*.

Spaced Repetition. The amount of time that elapses between successive repetitions of a given item strongly moderate memory for that item (see "Related Work", above). As such, we extracted a number of spacing-related features. To measure the temporal lag, and to capture the power law relationship between time and forgetting, we calculated (separately for each user) the log(*days*) that had elapsed between: (1) each token and its previous occurrence (*tokenLag1*), (2) each token's previous occurrence and its next most recent occurrence (*tokenLag2*), (3) each token's *stem* (e.g. *help,* for *helping*) and its previous occurrence (*stemLag1*), (4) each token's *stem*'s previous occurrence and its next most recent occurrence (*stemLag2*), (5) each token's combination of several morphological features (*number*, *person*, *tense*, *verbform*) and the previous occurrence of that particular combination (*morphoLag1*; to capture any possible spacing effect for verb conjugation skills) and (6) each token's combination of those same morphological features and their next most recent occurrence. Finally, (7) since some evidence suggests that the temporal gap between an item's first and second occurrence is particularly important for retention (Karpicke & Roediger, 2007), we also labeled each instance with the log(*days*) that elapsed between the first and second occurrence of the token's stem (*lagTr1Tr2*).

4.3 Categorical Features

Included in our classifier were a number of categorical features, each encoded as binary indicator variables distributed over a number of columns equal to the number of levels in the category. Importantly, our approach to modeling was constrained by limited computational power and memory, so we chose to include only categorical features with a relatively small number of levels, to reduce the dimensionality of the data. Those features were: *part of speech* (*pos;* 16 levels), *countries* (94 levels), *session* (3 levels), *format* (3 levels), and all of the morphological features available for each language (46 levels for learners of Spanish, 17 levels for learners of English, and 10 levels for learners of French). *Client* was also included, though we treated iOS and Android as equivalent, preserving only the distinction between web and mobile access to the Duolingo application (2 levels).

Notably, the above listing omits two of the categorical features we considered of greatest potential value in predicting early learner errors: *user* (223 levels for learners of Spanish, 179 levels for learners of English, and 216 levels for learners of French; 618 total) and *token* (2116 levels for learners of Spanish, 1615 in learners of English, 1682 in learners of French). Some users inevitably learn faster and make fewer errors than others, and some tokens are simply harder to learn on average. Instead of encoding these with dummy variables, we elected to replace the *user* feature with two continuous values, determined jointly by the user and the combination of the levels of the features *format*, *session*, and *client* for each instance: (1) the mean and (2) the variance of the error rate for that user under that combination of feature levels (*userMeanError*, *userVarError*, respectively), for a total of three values for each user. Similarly, we replaced the token feature with (1) the mean and (2) the variance of the error rate for each combination of the features *token*, *stem*, *format*, and *pos*, creating four values per token. This approach allowed us to substantially reduce demands on computational resources while simultaneously capturing much of the predictive power that fully encoding each user and token would have provided. The particular features used to create means within

user and token were chosen to maximize potential differences between accuracy in different modalities. Indeed, to foreshadow our results, these features each ranked among the most important for our random forest classifier.

4.4 Interactions

Several interactions between features were also coded into the model. Due to time constraints, only the following interactions were added: *stemLag1* x *stemLag2* and *stemLag1* x *stemLag2* x *lagTr1Tr2*, to capture spacing effects, *lagTr1Tr2* x *morphoComplexity* and *lagTr1Tr2* x *morphoLag1* to capture lag differences between morphological features, *format* x *prevFormat* to capture possible task switching effects, and *orthoNei* x *format* and *phonNei* x *format* and *format* x *client* to capture differences due to listening vs. typing, and finally *morphoComplexity* x *pos* as any complexity effect may be stronger nouns and verbs than function words.

5 Model

In order to focus on feature engineering, random forest techniques were chosen over gradient boosting, logistic regression or other classification techniques. The random forest classifier scales well to large datasets, is not particularly prone to overfitting problems, and requires less parameter tuning.

Random forest classifiers combine the outputs of multiple decision tree classifiers with random features taken in each decision in order to generate one final prediction (Breiman, 2001). Each decision-tree classifier split the data along some number of parameters (equal to the square root of the total number of features in this model) that fits a classifier. Each split of the data was again split along the other included parameters until the leaves of the tree contained only data points with the same label (i.e., only error or only no-error instances). For each learner population, we generated 1000 decision trees to generate predictions. Out-of-bag errors were used to estimate errors in training.

6 Results and Discussion

Figure 1 shows the top 20 importance scores for each language (out of an across-language total of 174 features or interactions). The importance

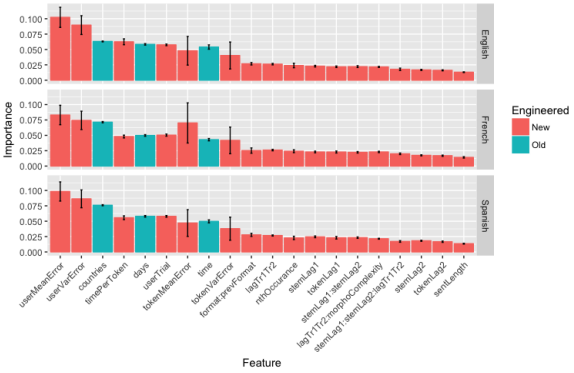

Figure 1: Top 20 importance features grouped by to-be-learned language. Error bars represent standard deviation of the importance of each feature across decision trees. For categorical features, the importances of each level, and their variances (to generate stand deviations), were summed to calculate the overall importance and variability in importance, respectively.

score of a random forest model conveys the predictive power of a given feature relative to the other predictors. Color depicts which features were engineered and which were provided in the raw data. Full importance values, for each language are listed in Appendix A, including the directionality of the relationship between each continuous feature and the error rate. For example, because *userMeanError* is higher on incorrect trials than correct trials, the directionality is considered positive.

The mean and variance in error rate for each user (*userMeanError* and *userVarError)* were the most important features, indicating that each user's history was strongly predictive of their performance at test, and that the variability within each user was nearly as predictive as the difference between users.

Countries, the third most important feature, may have ranked third in all three languages because the importance measure was calculated by summing over each feature level, possibly overstating the value of that feature in total. Nevertheless countries may represent user background information not given in the dataset including their previous language experience (as a Portuguese speaking user from Brazil may be learning Spanish via English, but would likely make different errors than an English monolingual from Canada).

The next most important generated feature was the average time spent on each token within an exercise (*timePerToken)*. This likely captures time

spent on each exercise better because it accounts for the length of the exercise at the token level.

Next is *userTrial*, which was calculatedly simply as which learning instance a given user is on. This likely captures the experience a user accumulates with the language and perhaps the app more generally.

Next of note is the mean and variance in error rate for each token, showing that each token has some properties that capture difficulty. This is especially true for learners of French, as the importance of *tokenMeanError* is ranked fourth in French as compared to eighth in both English and Spanish.

The interaction between format and previous format shows that there is some cost associated with task switching, perhaps to a slightly higher degree in English and Spanish, as this feature did not quite rank among the top ten in French, but was surpassed in that language by the lag between the first two occurrences of a token's stem.

Finally, the various lag features that reflect recent experience and many of their interactions comprise of the next most important features, indicating that spacing effects are generally predictive of errors of the overall model, the highest of these being the lag between the first two instances of a given token. This is an important and potentially useful feature. A measure of this lag is easy to calculate and necessarily occurs early in learning, making it useful in predictions that are memory intensive and catered to particular users or tokens.

Overall these features, and indeed many of the engineered features, improved the models over baseline, as seen in Table 1. This is particularly noteworthy considering user and token were removed in our model (and replaced with user- and token-level error rates), but were included in the SLAM baseline provided with the data. Indeed, the mean and variance across users and tokens account for ~25% of the importance across all languages. Though the importance of these features aggregate error rates in the training data, the metrics did not differ considerably when evaluated with the development data (AUROC = .824, .818, and .802 for English, French and Spanish respectively). This shows that aggregating is a feasible approach in cases where computational constraints prohibit the exact representation of important high dimensional categorical features. Notably, the within-user variability was an important feature in our model, but would not automatically be captured by dummy-coding user and token IDs across hundreds or thousands of instances. Thus, substantial computational savings can be achieved using low dimensional summary statistics where significant CPU time and memory resources would be required.

	AUROC	**F1**	**Log-loss**
SLAM English	.7730	.1899	.3580
English	.8286	.4242	.3191
SLAM French	.7707	.2814	.3952
French	.8228	.4416	.3561
SLAM Spanish	.7456	.1753	.3862
Spanish	.8027	.4353	.3571

Table 1: Final model outcomes in all three metrics as compared to baseline.

7 Future work

Due to the time-limited nature of this shared modeling task, considerable work remains to be done to both optimize the performance of this model and further understand the cognitive processes involved in early language learning.

To improve the model, we would first refine the relative importance of the current features, by performing ablation tests and model comparisons; some of the current features play little to no role in improving model performance. Furthermore, many interactions in the current feature space,

such as *userMeanAcc* x *tokenMeanAcc*, may be important predictors given each individual feature's importance, and that each user's previous language experience will impact the difficulty associated with any given token. The spacing effect might likewise interact with individual user and token related information.

There is additionally much work to be done in quantifying the benefit of using user- and token-level error rates as opposed to dummy-encoded variables. While these features are a memory and time sensitive solution, we have not yet explored how much model performance is affected by this change relative to a dummy-encoded solution, how much time is saved, and how much data is required to achieve this performance.

Our approach focused on linguistic and cognitive features that are known in their respective literatures to impact learning, and so the bulk of our efforts were devoted to feature engineering. Fu-

ture work will therefore dedicate more resources to model development. While in the present work only a random forest ensemble classifier was used to generate predictions, logistic regression, deep learning, and/or other modeling approaches may better suit this particular learning task, and should be thoroughly explored.

Finally, there are many more features than can be developed, including word embeddings of tokens and syntactic structure differences. Our work has scratched the surface of linguistic and cognitive theory that might be applied to modeling language learning, but the vast scientific literatures in those and other fields no doubt offer rich possibilities for new features. The relative contribution of all of these features and their interactions to machine learning models of error production is likely to greatly expand our knowledge of early second language learning.

8 Acknowledgements

We thank Ed Vul, Tim Sainburg, Vic Ferreira and the Language Production Lab for their feedback on this project.

References

Ben Ambridge, Anna L. Theakston, Elena V.m. Lieven, and Michael Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2), 174–193.

Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575-589.

Harry P. Bahrick and Elizabeth Phelphs. 1987. Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 344–349.

David A. Balota, Janet M. Duchek, and Ronda Paullin. 1989. Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3–9.

Leo Breiman. 2001. Random forests. *Machine Learning, 45(*1), 5-32.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Shana K. Carpenter. 2009. Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Shana K. Carpenter and Edward L. DeLosh. 2006. Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, 268-276.

Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. 2006. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380.

Nichlaos J. Cepeda, Edward Vul, Doug Rohrer, John T. Wixted, and Harold Pashler. 2008. Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095-1102.

William L. Cull. 2000. Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235.

Annette De Groot and Rineke Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1-56.

Ton Dijkstra, Mark Timmermans, and Herbert Schriefers. 2000. On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language*, *42*(4), 445-464.

John J. Donovan and David J. Radosevich. 1999. A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795-805.

Hermann Ebbinghaus. 1964. Memory: A contribution to experimental psychology (H.A. Ruger, C.E. Bussenius, & E. R. Hilgard, Trans.). New York, NY: Dover. (Original work published in 1885).

Jeffrey D. Karpicke and Henry L. Roediger. 2007. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology-Learning Memory and Cognition, 33*(4), 704-719.

Thomas K. Landauer and Robert A. Bjork. 1978. Optimum rehearsal patterns and name learning In M. M. Gruneberg, P. E. Morris, & R. N. Sykes 632). London: Academic Press.

Timothy D. Lee and Elizabeth D. Genovese (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, *59*(4), 277-287.

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, *7*(8), e43230.

Cornelius P. Rea and Vito Modigliani. 1985. The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research and Applications, 4*(1), 11-18.

B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. 2018. Second Language Acquisition Modeling. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL.

Michael T. Ullman and Jarrett T. Lovelett. 2016. Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *39*(1), 39-65.

Eleanor Vander Linde, Barbara A. Morrongiello, and Carolyn Rovee-Collier. 1985. Determinants of retention in 8-week-old infants. *Developmental Psychology*, *21*(4), 601–61.

Appendix A.

	English				French				Spanish			
Feature	**Import.**	**SD**	**Rank**	**Direc.**	**Import.**	**SD**	**Rank**	**Direc.**	**Import.**	**SD**	**Rank**	**Direc.**
Case	0.001	0.001	52						0.001	0.000	57	
client	0.004	0.001	44		0.005	0.001	41		0.005	0.001	38	
Concreteness (M)	0.007	0.001	31	+	0.011	0.004	26	-	0.010	0.001	25	+
Concreteness (SD)	0.007	0.001	34	-	0.009	0.003	31	+	0.010	0.001	27	+
countries	0.063	0.001	3		0.071	0.001	3		0.076	0.001	3	
days	0.058	0.001	5	+	0.050	0.001	6	+	0.058	0.001	4	+
Definite	0.000	0.000	55		0.001	0.000	54		0.001	0.000	55	
Degree	0.001	0.000	53						0.000	0.000	61	
dependencyEdgeHead	0.011	0.001	22	-	0.011	0.001	25	+	0.012	0.001	21	-
editDistance	0.007	0.001	32	-	0.010	0.003	29	-	0.010	0.001	26	+
EngPhos	0.005	0.002	43	-	0.006	0.002	38	-	0.002	0.000	50	-
Foreign									0.000	0.000	63	
format	0.006	0.006	36		0.008	0.009	32		0.008	0.008	30	
format:client	0.010	0.006	24		0.012	0.009	22		0.012	0.007	22	
format:prevFormat	0.027	0.002	10		0.025	0.004	11		0.028	0.003	10	
Gender	0.000	0.000	54		0.004	0.001	47		0.004	0.000	42	
Homograph	0.002	0.001	49	-	0.002	0.001	51	-	0.002	0.001	49	-
lagTr1Tr2	0.026	0.001	11	+	0.026	0.001	10	+	0.027	0.001	11	+
lagTr1Tr2:morphoComplex	0.022	0.001	16	+	0.023	0.001	13	+	0.022	0.001	16	+
logEngPhoNeiFreq	0.005	0.001	42	-	0.006	0.002	36	-	0.002	0.000	52	+
logOrthoNeiFreq	0.007	0.001	35	-	0.007	0.002	33	-	0.005	0.001	39	-
logPhonNeiFreq	0.006	0.001	37	-	0.006	0.001	37	-	0.005	0.001	40	-
logWordFreq	0.008	0.002	28	-	0.009	0.003	30	-	0.005	0.001	35	-
Mood	0.002	0.000	50		0.001	0.000	56		0.001	0.000	56	
morphoComplex	0.002	0.000	48	-	0.003	0.002	50	-	0.003	0.000	46	-
morphoComplex:pos	0.006	0.001	38		0.007	0.002	35		0.006	0.001	33	
morphoLag1	0.008	0.000	30	+	0.005	0.000	42	+	0.005	0.000	36	+
morphoLag1:morphoComplex	0.008	0.000	27	+	0.005	0.000	43	+	0.005	0.000	37	+
morphoLag2	0.010	0.000	23	+	0.006	0.000	39	+	0.005	0.000	34	+
nthOccurance	0.024	0.004	12	-	0.024	0.002	12	-	0.023	0.003	15	-
Number	0.002	0.000	47		0.005	0.002	45		0.004	0.001	41	
NumType	0.000	0.000	56						0.000	0.000	59	
orthoLength	0.004	0.002	45	+	0.005	0.001	46	+	0.003	0.001	47	+
OrthoNei	0.005	0.001	39	-	0.006	0.002	40	-	0.004	0.001	43	-
OrthoNei:format	0.010	0.004	26		0.014	0.008	21		0.008	0.003	31	
Person	0.001	0.001	51	+	0.004	0.001	48		0.003	0.000	48	
phoLength	0.004	0.001	46	+	0.004	0.001	49	+	0.003	0.001	45	+
PhonNei	0.005	0.001	41	-	0.005	0.002	44	-	0.003	0.000	44	-
PhonNei:format	0.008	0.004	29		0.011	0.007	24		0.007	0.003	32	

Polite									0.000	0.000	64	
pos	0.007	0.001	33		0.012	0.004	23		0.009	0.000	29	
Poss									0.000	0.000	58	
PrepCase									0.000	0.000	60	
PronType					0.002	0.001	52		0.002	0.000	51	
Reflex									0.000	0.000	62	
sentLength	0.013	0.001	20	+	0.014	0.001	20	+	0.013	0.001	20	+
session	0.010	0.001	25		0.011	0.001	28		0.011	0.001	23	
stemLag1	0.023	0.001	13	+	0.023	0.001	14	+	0.025	0.001	12	+
stemLag1:stemLag2	0.022	0.001	14	-	0.022	0.001	16	-	0.023	0.001	13	-
stemLag1:stemLag2:lagTr1Tr2	0.018	0.002	17	+	0.020	0.001	17	+	0.017	0.001	18	-
stemLag2	0.017	0.001	18	+	0.018	0.001	18	+	0.018	0.001	17	+
Tense					0.001	0.000	55		0.001	0.000	54	
time	0.054	0.003	7	+	0.043	0.002	8	+	0.050	0.002	7	+
timePerToken	0.062	0.005	4	+	0.048	0.002	7	+	0.056	0.003	6	+
tokenIndex	0.012	0.001	21	+	0.011	0.001	27	+	0.011	0.001	24	+
tokenLag1	0.022	0.001	15	+	0.023	0.002	15	+	0.023	0.002	14	+
tokenLag2	0.016	0.001	19	+	0.017	0.001	19	+	0.017	0.001	19	+
tokenMeanError	0.048	0.023	8	+	0.070	0.033	4	+	0.047	0.022	8	+
tokenVarError	0.040	0.022	9	+	0.042	0.022	9	+	0.038	0.019	9	+
userMeanError	0.102	0.016	1	+	0.083	0.016	1	+	0.098	0.016	1	+
userTrial	0.058	0.001	6	+	0.050	0.002	5	+	0.058	0.001	5	+
userVarError	0.090	0.015	2	+	0.074	0.015	2	+	0.086	0.015	2	+
VerbForm					0.001	0.000	53		0.001	0.000	53	
wordLength	0.005	0.002	40	+	0.007	0.003	34	+	0.009	0.002	28	+