Nosov Andrei Valerievich

Perm National Research Polytechnic University,
29, Komsomol'sky av., Perm, 614990, Russia;
andrey.nosoff@gmail.com

Statistical analysis of near-synonymous words *list* and *catalog* in R

For citation: Nosov A. V. Statistical analysis of near-synonymous words *list* and *catalog* in R. *Vestnik of Saint Petersburg University. Language and Literature*, 2018, vol. 15, issue 3, pp. 453–464. https://doi.org/10.21638/spbu09.2018.310

In this article, we present the results of the regression analysis of near-synonymous words *list* and *catalog*. The purpose of the case study is allocation of the most objective variant by modeling the grammatical interactions that make impact on updating of the considered words. Determination of *list* and *catalog* as objective and independent lexical units is performed within the system of distinctions and oppositions. By the probabilistic distribution, we allocate two most frequent interactions. The comparison of average values does not reveal regularly all aspects of the studied phenomenon (i.e. average values of models can be statistically identical). Therefore, we compare the models with predictors PRE.MOD and GENITIVE MEAN with the model without interactions to show distinction between them at the level of dispersion. Hence, three statistical hypotheses are compared in pairs. The main says that dispersions of three considered models are statistically equal and the alternative affirms that they are different. Model assessment without interactions is a predictive *logit* of *list*. Coefficients of logistic regression reflect the probability of changes within interactions. At the stage of normalization, we apply the model of the binary choice Hosmer—Lemeshow. Based on the obtained results we decide whether it is necessary further normalization or not. We define also the presence/absence of correlated samples among the considered predictors by *lrm* function, which determines reliability of the model and allows receiving confidential intervals of coefficients. This approach reflects novelty of work and allows revealing the factors defining the choice of one or another concept proceeding from objective semantic criteria. Interactions are considered at four levels: academic, spoken, fiction and news. Results of research allow to complete the content of the words *list* and *catalog* and to present their dynamics.

Keywords: computational linguistic, logistic regression, comparative analysis, semantics, synonym, *list*, *catalog*.

1. Introduction

In order to define the place of our approach in the modern statistical research we note the complexity in application of near-synonymous words that motivates researchers in the field of lexical semantics to focus their attention on this phenomenon. We pay attention on varying degrees of loose synonymy with the aim to differentiate the words *list* and *catalog*. Hence, we refer to G. A. Miller & G. Ch. Walter [Miller, Walter 1991] who identify not only a significant conjunction in meaning between two words, but also some contextual representation. This idea is supported by G. Leitner [Leitner 1993] in his work

on the meaning of the words in the 'real English' and developed by S. Hunston adopting corpus-based analyses of words within this problem [Hunston 2002]. Also, S. Th. Gries and others exemplified corpora based on naturally-occurring language samples [Gries 2001: 4; Gibbs 2006; Geeraerts 2010].

Nowadays the widespread application of R tools and the development of corpus semantics create a strong foundation in the study of near-synonymous words. Following the corpora K. W. Church, W. Gale, P. Hanks, D. Hindle and R. Moon [Church et al. 1994] study the synonyms *ask for*, *request* and *demand*. N. Levshina, D. Geeraerts and D. Speelman [Levshina et al. 2014] investigate the difference between the Dutch causative verbs *doen* and *laten*, G. Gilquin [Gilquin 2003] analyzes verbs *get* and *have*, S. Th. Gries [Gries 2001] compares English adjectives ending in *-ic* or *-ical*. D. Glynn [Glynn 2010] quantifies the similarity between *hassle*, *bother* and *annoy*. S. Phoocharoensil [Phoocharoensil 2010] examines *ask*, *beg*, *plead*, *request*, and *appeal*, concentrating on their lexical, syntactic, and stylistic information. S. Th. Gries and N. Otani [Gries, Otani 2010] test the near-synonyms *big*, *great* and *large* in R.

Mostly they support the behavioral approach to gain round insights in the analysis of semantic differences of near-synonymous words. In this case study, we will follow them to a certain extent.

2. Background[1]

Firstly, we allocate criteria for synonyms. In the Modern Linguistics, we can differentiate synonyms by collocation analysis, definition of context register (style), grammatical pattern differentiation dialect usage referential analysis (connotations).

A glimpse at English dictionary definitions detects rather a few similarities in meaning between *list*, *catalog*, *register*, *schedule*, *nomenclature*, *roll* and *inventory*. In fact, in certain cases the definitions of these lexical items seem to be circular. Considering these notions within a computational approach and frequencies extracted from COCA (Corpus of Contemporary American English) reduces the collection to two most frequent samples. Hence, the most accurately allocated opposition is *list* (55 111 occurrences in COCA) and *catalog* (1813 occurrences in COCA)[2].

Once we consider the definitions of *list* taken from several dictionaries [The New Collins Thesaurus 1984; The New Shorter Oxford English Dictionary 1993; Collins English Dictionary 1994; Random House Webster's College Dictionary 1995; Longman Dictionary 2009] it becomes evident that *register*, *schedule*, *nomenclature*, *inventory*, *roll* almost infallibly occur as one of the variations of *list*; at the same time *catalog* also has a strong resemblance to *list*. Thus, we put a task to find out the main lines of these two concepts.

Firstly, we give their definitions below:

List [The Free Dictionary 2003–2018]

1. A series of names or other items written or printed together in a meaningful grouping or sequence so as to constitute a record.
2. Computing a linearly ordered data structure.
3. A database containing an ordered array of items (names or topics).

[1] Based on: [Nosov 2016].

[2] The examples are from Davies M. The Corpus of Contemporary American English (COCA): 425 million words, 1990 [Americancorpus.org S. a.].

4. Item, point.
5. An alphabetical index of names and topics along with page numbers where they are discussed.

Catalog [The Free Dictionary 2003–2018]

1. A list or itemized display, as of titles, course offerings, or articles for exhibition or sale, usually including descriptive information or illustrations, a publication, such as a book or pamphlet, containing such a list or display.
2. An enumeration.
3. A card of the contents of a library or a group of libraries, arranged according to any of various systems.
4. Any record.
5. A written work or composition that has been published (printed on pages bound together).
6. A complete list of things; usually arranged systematically.
7. A series, as of names or words, printed or written down.

By comparison of these dictionary definitions:

1. We note that the words *list* and *catalog* in modern English are poly-semantic (in the analysis I take into consideration only nucleus dictionary definitions which allow me to define their values in modern English; we do not consider obsolete definitions or others used for special purposes).
2. We appreciate the componential structure containing rather similar integrated and differential components. The Presence of integrated components in the structure of concepts under consideration is reflected by five groups of semantic elements:
   1) *series, set, sequence, grouping, array*;
   2) *things, names, numbers, items, members, words, topics, files, point*;
   3) *data structure, database, record*;
   4) *written, printed, imagined, ordered, contributed and stored*;
   5) *computing, containing, listing.*

These two near-synonymous words present difficulties in the semantic level. Our aim is to define the most appropriate for its usage in the further investigations on natural language. These units can be defined by distinctions and oppositions as well as by the analysis of grammatical constructions. Thus, we seek to identify the factors that determine the choice between *list* and *catalog* by objective and semantic criteria. The analysis is based on multiple logistic regression analysis developed in the program R, which «seems to have become the *de facto* standard tool in many areas of linguistics especially corpus based and computational studies» [Levshina 2015: 21]. In so doing we bring to the focus the testing of the hypotheses below:

1. The scope of the *catalog* is narrower than the *list* one;
2. The use of *catalog* is appropriate in more *academic* context whereas *list* is introduced in more universal way and equally represented in all mentioned registers;
3. The academic register like equally appropriate for both *list* and *catalog*;
4. *List* tends more than *catalog* to expanding semantics and not to narrowing;
5. *List* is more pre-modified than *catalog*;

6. *List* is more post-modified than *catalog* in all mentioned registers;
7. Both *list* and *catalog* being self-determinate are not frequently represented within genitive constructions;
8. *List* in genitive constructions favor less written registers (newspaper, fiction, and academic).

For the analysis and verification of hypotheses considered above, we use the **logit model**. In statistics, the logistic regression is a model used to predict the probability of events adapting the data to the logistic curve. Usually predictive variables either numeric or categorical can be used.

3. Statistical analysis of near-synonymous words *list* and *catalog*

For further analysis, we will use packages 'effects' and 'logreg', which can be freely downloaded from the online public platform CRAN [CRAN S. a.]. We activate them with commands:

library (effects)
library (rms)

These libraries enable to use not only *glm()* function [Cross-validated 2017], but activate *lrm()*, which allows us to present more clearly and accurately the results of the logit regression.

Thus, we will consider three models, one with main effects and the others with PRE. MOD and GENITIVE.MEAN interactions. The summaries of models are presented in Table 1, which is modeled after the one in «Corpus methods for semantics: Quantitative studies in polysemy and synonymy» [Levshina et al. 2014].

Table 1. **Summary Statistics**

Summary statistics	Model		
	with main effects only	with main effects and GENITIVE.MEAN interaction	with main effects and POST.MOD interaction
Number of observations	1600 (of which 62 catalog and 1538 list)		
Null deviance	524.64 (on 1599 df)	524.64 (on 1599 df)	524.64 (on 1599 df)
Residual deviance	486.29 (on 1592 df) [AIC is 502.29]	482.44 (on 1589 df) [AIC is 504.44]	474.72 (on 1589 df) [AIC is 496.72]
Model chi-squared	38.35 (on 7 df)	42.20 (on 10 df)	49.92 (on 10 df)
p-value (chi-squared)	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
R-squared	0.085	0.093	0.110
C (area under ROC curve)	0.714	0.723	0.761

The null deviance fixes the difference between the intercept model and the input data, and the deviance shows the difference between the fitted model and the input. That is to say, if we have a small deviance the model fits better the input data. In Table 1 we see that

the model with main effects and GENITIVE.MEAN interaction (RD 482.44) is better than the model with main effects only (RD 486.29). However, the model with main effects and POST.MOD interaction works much better (RD 474.72).

The degrees of freedom of the null deviance (sample size minus one) and the degrees of freedom of the deviance (sample size minus one minus number of regressors) are additional pieces of information that will be used in a log likelihood ratio test (LRT) that tests whether the reduction of deviance when moving from the intercept only model to the fitted model is significant or not (in other words, whether the difference between null deviance and deviance is significant). The AIC: value is a corrected version of the deviance that has the additional benefit that it can be compared across models (for the same data set and response variable) with different numbers of predictors. Finally the **Number of Fisher scoring iterations**: the section shows us how many iterations we need to converge upon the reported estimates for intercept and slope.

From the Table 1, we can read that in the model with main effects only, the p-value is less than 0.05 ($p < 0.0001$). This fact and associated **Chi-square** (38.35 with 7 degrees of freedom) give proof that the model explains variation better than the intercept only model, i. e. the model with no predictors at all. The simple proportion of success or correct predictions in the model is 0.9650 while the baseline is at 0.96125. The calculation of this baseline is simple: our intercept only model simply predicts all cases to have a probability of success of 1538/1600 = 96 % (which is the overall proportion of success in the data set). Since 96 % is more that 50 %, the intercept only model would, therefore (according to our simple classification rule) always predict success, which would be a correct prediction in 96 % of the cases. Based on it we can state that simple classificatory success for our data goes up from the baseline of 96 % for the intercept model to 99 % for the fitted model.

The estimate for the intercept is 3.0556, and appears to indicate that concept *list* without predictors has an average logit of 3.0556 (corresponds to odds of 36 to 1 and a proportion of 97 %).

The estimate for the intercept is 3.0556, which is the predictive logit [Minitab Inc. 2010] of *list* for the case with all the reference levels, that is, **academic register, no possessive meaning, post-modification, no genitive meaning, pre-modification**. In other words, the positive value of the intercept means that the so-called failure response, i.e. the variant *catalog* is less popular in the case of the academic register, no possessive meaning, no post-modification, no genitive meaning, pre-modification. If we turn this logit into odds and then in probability what we get are odds of 36 to 1 and a proportion of 97 %, which is the predicted probability of *list* in the case of academic register, no possessive meaning, no post-modification, no genitive meaning, pre-modification. According to the significance test, it is different from a logit of zero (i. e. at odds of 1 to 1 and a proportion of 50 %), hence the model is suitable.

The coefficients of the logistic regression reflect the probability changes of the concept according to the corresponding predictor. In any case that contains register = academic and that we want to switch from, to, say, register = fiction, the logit of *list*, and hence its probability increases by 0.6855[3]. Whatever the other circumstances, there is always the same effect of register = fiction.

[3] When the logit goes up, the probability goes up as well and if the logit decreases so does the probability [Speelman 2014].

Then, it is possible to go from register = academic to register = news, and the logit in this case, would increase by 1.2747, or if we switch from register = academic to register = spoken, logit again increases, but this time by 2.3943. If there is a switch from PRE.MOD = no to PRE.MOD = yes the logit goes up by 0.1757. In the case of POST.MOD = no to POST.MOD = yes the logit goes down by –0.6892, and the same picture we observe in the case of GENITIVE.MEANING = no to GENITIVE.MEANING = yes when the logit reduce to –0.7600. Finally, the logit increases by 0.2572 if there is a switch from POSS. MEAN = no to POSS.MEAN = yes. It might be interesting to note that the model assigned the highest probability to *list* (99,4 %) in the case of register = spoken, POSS.MEANING = no, GENITIVE.MEAN = no, POST.MOD = yes, PRE.MOD = yes[4]. Figure 1 (below) can help us visualize the effects of all predictors according to this model.

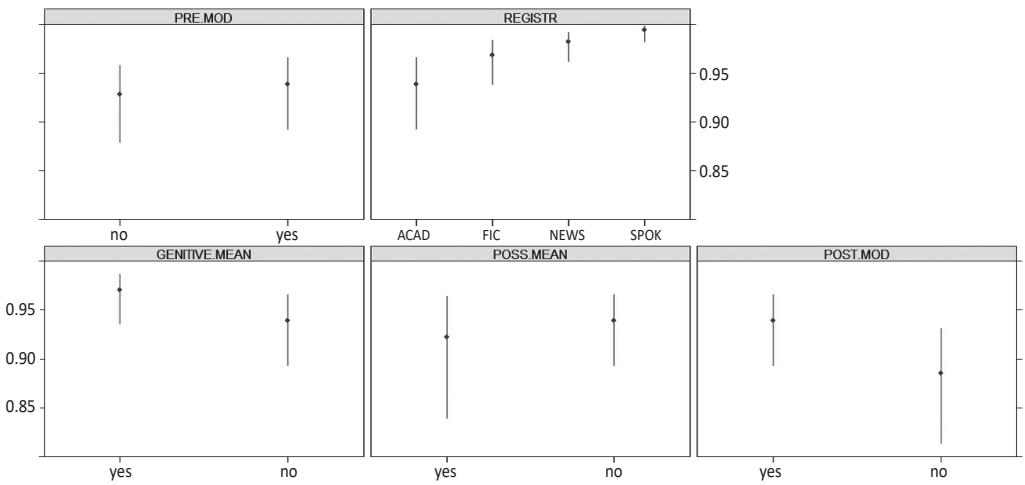

Figure 1. Predicted probabilities of *list*

Testing to see whether all the predictors are significant revealed that they were, with register being the most significant one and pre.mod the least significant of the five. When compared to other predictors, pre.mod has a small effect size in terms of unexplained variation (LRT = 0.396).

Acting on the assumption that emerged after inspecting the data that there might be some interaction between a register and genitive meaning, I decided to run a second model, which would consider this interaction. Simply put, the interaction occurs whenever the effect of one predictor is conditioned by another predictor. It was expected that genitive meaning would mostly favor *list* in the academic register.

First, we can look at the frequency table of the register by genitive meaning by the variant in Figure 2. What we can see is that there are more observations of *list* than *catalog* in the data. *List* is popular across all registers. It seems that compared to *catalog* genitive

[4] Regarding the confidence interval around the estimates, there is 95 % confidence about intercept logit being between 0.8928548 and 0.9656735; REGISTR = FIC between 0.9382686 and 0.9838080; REGISTR = NEWS between 0.9616903 and 0.9917038; REGISTR = SPOK between 0.9819396 and 0.9980730; POSS. MEAN = no between 0.8392593 and 0.9640849; PRE.MOD = yes between 0.8793471 and 0.9576874; POST. MOD = yes between 0.8928548 and 0.9656735; GENITIVE.MEAN = no between 0.9360397 and 0.9865293.

meaning favors it more. The question is whether the distance is considerable or marginal [Speelman 2014]. However, since this is just a frequency table that takes into account register and genitive meaning and not the other predictors, i. e. post.mod, which was also proven to have an influence, it is not very good at telling us where the effect comes from, which the model with interaction and all the other predictors should be able to.

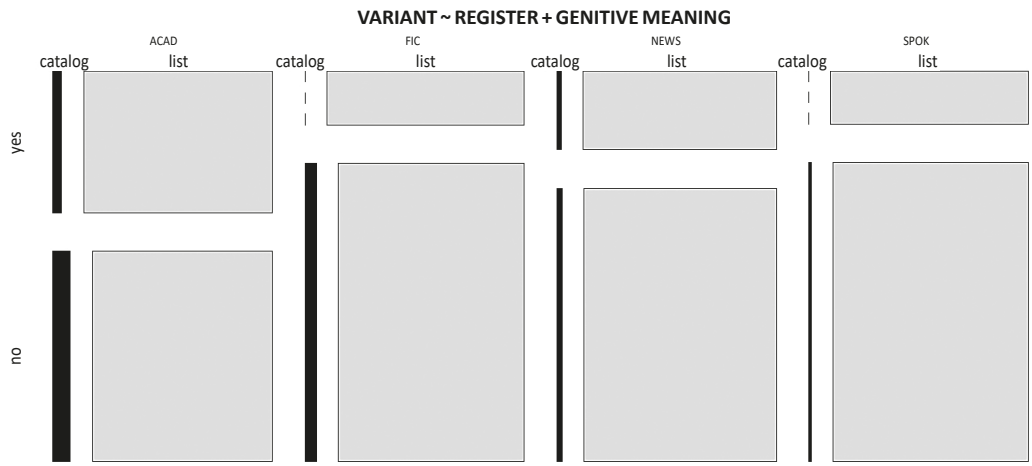

Figure 2. Mosaic frequency plot of register + genitive meaning

What can be read in the summary statistics Table 1 (above) is that the model with interaction is significant (p-value < 0.0001). The estimate for the intercept is 2.9684, which again is the predictive logit for the case with all the reference levels, academic register, no possessive meaning, no post-modification, no genitive meaning, pre-modification. The test to see whether the interaction was significant showed that it was not, with p-value < 0.0001. The amount of unexplained variation in the model with all predictors including the interaction is slightly bigger (AIC is 504.44) than in the model without interaction (AIC is 502.29).

The results of logistic regression showed here that no genitive meaning is appropriate to *list* throughout the registers, but mostly in spoken register (84 %) a bit less, but still considerable in fiction register (79,8 %), and finally, in news (75,8 %) and academic register (54,5 %), respectively.

The post.mod predictor is also significant, 3.37e-10 ***, in fact, much more so than the previous interaction (7.62e-08 ***). However, since dropping the interaction from the model would lead to the significant increase of unexplained variation, it was decided to keep the previous one in the model. All indexes (AIC is 496.72; R squared = 0.110; C (area under ROC curve) = 0.761) showed that this model works best.

In the mosaic plot, we see that *list* is more post-modified than the *catalog* in all the registers.

After the analysis with post-modification interaction, we can conclude that in support of the data presented in Table 1 post-modification is appropriate to *list* throughout all the registers, except spoken one (58,2 % of no post-modification cases). Mostly this effect is observed in the academic register (66 %), in the register of fiction, there are only 49 % instances of *list* with post-modification, while in news it is 60 %.

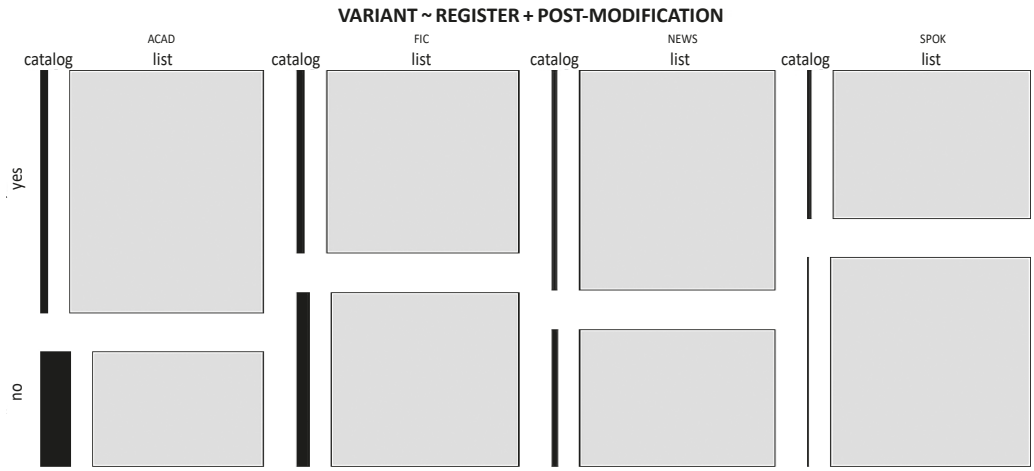

Figure 3. Mosaic frequency plot of register and post-modification

Further, we perform the Hosmer–Lemeshow goodness of fit test. [Shah, Barnwell 2003; Hosmer, Lemeshow 1989].

A significant result in that test would indicate a significant lack of fit. We do not obtain a significant lack of fit here (p = 0.05970231), but the rather low p-value does indicate that the fit is a bit far from good.

Another thing that should be tested routinely is whether there are no correlation patterns among the predictors that are so outspoken that this makes the model less reliable. We test it running on the *lrm* model without interactions. Thus we get the confidence intervals of calculated coefficients. See Table 2.

Table 2. **Coefficients of 'Odds Ratio' via Function *exp(coef(d.glm)***

REGISTR = FIC	REGISTR = NEWS	REGISTR = SPOK	POSS.MEAN = no	PRE.MOD = yes	POST.MOD = no	GENITIVE. MEAN = no
1.311409	1.203975	1.117437	1.140100	1.126551	1.096815	1.114420

In the results, we do not have values higher than **four** that means we should not remove any variable from the model and test it again.

And last but not least, considering the odds are 36 (i. e. 36 to 1) at the slope of intercept we request the 95 % confidence intervals for the estimates as follows. See Table 3.

Table 3. **Confidence Intervals for Model Parameters via Function *confint(d.glm)***

Model Parameters	2.5 %	97.5 %
(Intercept)	2.08945336	4.12170716
REGISTRFIC	0.06184795	1.32701710
REGISTRNEWS	0.54906435	2.07802479
REGISTRSPOK	1.42015910	3.63509215
POSS.MEAN = no	–0.47672834	0.92817534
PRE.MOD = yes	–0.37413155	0.72492486
POST.MOD = no	–1.24227974	– 0.14884359
GENITIVE.MEAN = no	–1.58105048	– 0.03975219

4. Results

The aim of this case study was to find the most appropriate between *list* and *catalog* to use in the further investigations. We consider them as near-synonymous words seeking to identify some of the factors that determine our choice.

Realization of this purpose demanded to solve the following problems:

1) to present semantic classification of these nouns which are used in the form of plural;
2) to reveal the semantic conditions promoting a specification of *list* and *catalog* and their use in the form of plural, having shown primacy of semantic factors in relation to the formal;
3) to define a circle of contexts in which *list* and *catalog* are most often used;
4) to track interaction of semantic quantified POST.MOD and GENITIVE.MEAN the factors reflecting the functioning of these words.

The method that we used for this purpose was multiple logistic regression analysis. The focus of the investigation was on the identification of the contents and dynamics of the concept verbalized by *list* and *catalog* across four registers of the language (academic, spoken, fiction, newspaper), the semantic features of possessive meaning, genitive meaning, plural modification, post-modification, and pre-modification.

5. Conclusions

1. The investigation showed that the semantic field of *catalog* is much narrower than those of *list*.

2. We find *catalog* mostly in specialized, professional and academic contexts whereas *list* is inherent for a freer contextual environment of a spoken language. In spite of the fact that the percent of use of *list* at spoken register is rather high (11 181 occurrences), the newspaper register (17 812 occurrences) remains still the most peculiar for it. Hence, the hypothesis that the word *list* is inherent for freer contextual environment of a spoken language came true partly.

3. In the dataset of 1600 observations, there are 281 cases of POSS.MEAN = yes and 1319 cases of POSS.MEAN = no. It should be recognized that the present case study did not look at the possessive meaning, which could have certainly added another dimension to the investigation of differences between *list* and *catalog*. Furthermore, it might be the case that some more fine-grained distinctions emerge from a much larger dataset. Therefore, it would not be wrong to take this case study as a stepping-stone towards a more comprehensive research.

4. There are only six cases of plural modification in a random sample taken from the corpus COCA that allows concluding that the plural form with resumptive and separative semantic features is marginally represented or mostly omitted within *list* and *catalog*. The assumption that *list* is not frequently represented within genitive meaning could be said to be confirmed in the dataset if we consider its distribution in the academic register compared to the other registers, especially spoken (84 % no genitive meaning). However, it can also be contended that it is indeed a more general word than *catalog* bearing in mind its overall frequency across the registers of the English language. Another assumption that genitive constructions favor written registers seems also to be borne out. This does not

come as a surprise since even some dictionaries, most notably, the "New Shorter Oxford English Dictionary…" [The New Shorter Oxford English Dictionary 1993] draw the users' attention to this reading of the word as usually accompanied by genitive constructions. However, taking it a step further and observing the semantics of the word from the perspective of the conceptual metaphor theory, it was possible to adopt a broader take on this particular meaning of *list*.

5. Pre-modification seems to be irrelevant with 868 cases of PRE.MOD = yes and 732 cases of PRE.MOD = no.

6. Post-modification of a lexical item might come in different forms, but in this case study, its general manifestation was taken into account. The prediction that post-modification favors *list* in all the registers and mostly in academic was confirmed in my dataset. Additionally taking the interaction between register and post-modification into consideration revealed that even though post-modification can be expected after *list* in all registers, this most frequently happens to be the case in academic English, which is excellent since pre-modifiers and post-modifiers are expected to be "rare in conversation, and very common in informational writing" [Longman Grammar 1999]. Finally, the lowest realization of *list* is indeed in the register of spoken English right after the register of fiction English.

Based on the obtained results within statistical analysis, it is possible to draw a conclusion that the concept of *list* favors the application in further research.

Dictionaries and guides

Americancorpus.org S. a. — *Americancorpus.org*. S. a. URL: http://www.americancorpus.org (accessed date: 29.05.2017).

Collins English Dictionary 1994 — *Collins English Dictionary*. Glasgow: HarperCollins, 1994, 1791 p.

CRAN S. a. — "CRAN". *The Comprehensive R Archive Network*. S. a. URL: https://cran.r-project.org/web/packages/effects/index.html. (accessed date: 29.05.2017).

Cross-validated 2017 — "Cross-validated". S*tack Exchange Inc. Educational portal: Interpretation of Multiple Logistic Regression with Interactions in R*. 2017. URL: http://stats.stackexchange.com/questions/115188/interpretation-of-multiple-logistic-regression-with-interactions-in-r (accessed date: 29.05.2017).

Longman Dictionary 2009 — *Longman Dictionary of Contemporary English*. 5th ed. Harlow: Pearson Longman, 2009. 2081 p.

Longman grammar 1999 — *Longman Grammar of Spoken and Written English*. Biber D., Johansson S., Leech G. et al. (eds.). Harlow: Longman, 1999, 1204 p.

Random House Webster's College Dictionary 1995 — *Random House Webster's College Dictionary*. New York: Random House, 1995, 1567 p.

The Free Dictionary 2003–2018 — *The Free Dictionary by Farlex*. 2003–2018. URL: www.thefreedictionary.com (accessed date: 29.05.2017).

The New Collins Thesaurus 1984 — *The New Collins Thesaurus*. McLeod W. T. (ed.). London: Collins, 1984, 759 p.

The New Shorter Oxford English Dictionary 1993 — The *New Shorter Oxford English Dictionary on Historical Principles*: in 2 vols. Brown L. (ed.). Oxford: Clarendon, 1993, 3801 p.

References

Church et al. 1994 — Church K. W., Gale W., Hanks P., Hindle D., Moon R. "Lexical substitutability". *Computational Approaches to the Lexicon*. Atkins B. T. S., Zampolli A. (eds.). Oxford: Oxford University Press, 1994, pp. 153–177.

Geeraerts 2010 — Geeraerts D. *Theories of Lexical Semantics*. Oxford: Oxford University Press, 2010, 341 p.
Gibbs 2006 — Gibbs R. W. "Metaphor Interpretation as Embodied Simulation". *Mind & Language*. 21 (3), 2006: 434–458.
Gilquin 2003 — Gilquin G. "Causative 'Get' and 'Have': So Close, So Different". *Journal of English Linguistics*. 31 (2), 2003: 125–148.
Glynn 2010 — Glynn D. "Synonymy, Lexical Fields, and Grammatical Constructions. Developing Usage-based Methodology for Cognitive Semantics". *Cognitive Foundations of Linguistic Usage Patterns*. Schmid H.-J., Handl S. (eds.). [Berlin; New York]: De Gruyter Mouton, 2010, pp. 89–118.
Gries 2001 — Gries S. Th. "A Corpus-linguistic Analysis of -ic and -ical Adjectives". *ICAME Journal*. 25, 2001: 65–108.
Gries, Otani 2010 — Gries S. Th., Otani N. "Behavioral Profiles: A Corpus-based Perspective on Synonymy and Antonymy". *ICAME Journal*. 34, 2010: 121–150.
Hosmer, Lemeshow 1989 — Hosmer D. W., Lemeshow S. *Applied Logistic Regression*. New York: Wiley, 1989, XIII, 307 p.
Hunston 2002 — Hunston S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002, 241 p.
Leitner 1993 — Leitner G. "Where to 'Begin' or 'Start'? Aspectual Verbs in Dictionarie". *Data, Description, Discourse: Papers on the English Language in Honour of J. McH Sinclair on His 60th Birthday*. Hoey M. (ed.). London: Harper Collins, 1993, pp. 50–63.
Levshina 2015 — Levshina N. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam; Philadelphia: John Benjamins, 2015, 443 p.
Levshina et al. 2014 — Levshina N., Geeraerts D., Speelman D. "Dutch Causative Constructions with Doen and Laten: Quantification of Meaning and Meaning of Quantification". *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Glynn D., Robinson J. (ed.). Amsterdam: John Benjamins, 2014, pp. 205–221.
Miller, Walter 1991 — Miller G. A., Walter G. Ch. "Contextual Correlates of Semantic Similarity". *Language and Cognitive Processes*. 6 (1), 1991: 1–28.
Minitab Inc. 2010 — "Minitab Inc.". *Softline Ltd. Educational portal*. 2010. URL: http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/regression-models/what-are-response-and-predictor-variables/ (accessed date: 29.05.2017).
Nosov 2016 — Nosov A. V. "Lingvisticheskie parametry kontseptov «list» i «catalog»: Variant obrabotki iazyka dlia komp'iuternykh system [Linguistic Parameters of the Concepts "LIST" and "CATALOG": Language Processing Version for Computer Systems]". *Vestnik Permskogo un-ta: Rossiskaia i zarubezhnaia filologiia* [Bulletin of Perm University: Russian and Foreign Philology]". 4 (36), 2016: 75–82. (In Russian)
Phoocharoensil 2010 — Phoocharoensil S. A. "Corpus-Based Study of English Synonyms". *International Journal of Arts and Sciences*. 3 (10), 2010: 227–245.
Shah, Barnwell 2003 — Shah B. V., Barnwell B. G. "Hosmer-Lemeshow Goodness of Fit Test for Survey Data Research". *2003 ASA Proceedings: Papers Presented at the Annual Meeting of the American Statistical Association: Joint Statistical Meetings, San Francisco, California, August 3–7, 2003, and Other ASA-sponsored Conferences*. S. l.: American Statistical Association, 2003, pp. 3778–3781.
Speelman 2014 — Speelman D. "Logistic Regression: A Confirmatory Technique for Comparisons in Corpus Linguistics". *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Glynn D., Robinson J. (eds.). Amsterdam: John Benjamins, 2014, pp. 487–533.

Received: November 9, 2016
Accepted: August 9, 2017

Носов Андрей Валерьевич

Пермский национальный исследовательский политехнический университет,
Россия, 614990, Пермь, Комсомольский пр., 29;
andrey.nosoff@gmail.com

Статистический анализ близких по значению слов *list* и *catalog* в программе R

Для цитирования: *Nosov A. V.* Statistical analysis of near-synonymous words *list* and *catalog* in R // Вестник Санкт-Петербургского университета. Язык и литература. 2018. Т. 15. Вып. 3. С. 453–464. https://doi.org/10.21638/spbu09.2018.310

В данной статье приводятся результаты регрессионного анализа двух близких по значению слов *list* и *catalog*. Целью анализа является выделение наиболее объективного варианта на основе моделирования грамматических интеракций, оказывающих влияние на актуализацию рассматриваемых слов в языке. Определение *list* и *catalog* в качестве объективных и независимых лексических единиц осуществляется в системе различий и противопоставлений. На начальном этапе согласно вероятностному распределению выделяются два наиболее частотных типа интеракций при употреблении слов *list* и *catalog*. Затем модели с предикторами PRE.MOD и GENITIVE.MEAN сопоставляются с моделью без интеракций, что продиктовано необходимостью показать различие между моделями на уровне дисперсии, так как сопоставление средних значений не всегда выявляет все аспекты изучаемого явления (ср.: средние значения моделей могут быть статистически одинаковыми). Таким образом, три статистические гипотезы сопоставляются попарно. Основная, нулевая гипотеза состоит в том, что дисперсии трех рассматриваемых моделей статистически одинаковы, и альтернативная — в том, что эти дисперсии статистически различны. Оценка модели без интеракций является предсказательным логитом *list* для вышеуказанных уровней отсчета. Коэффициенты логистической регрессии отражают вероятность изменений при взаимодействии с тем или иным предиктором. На этапе нормализации применяется модель бинарного выбора Хосмера—Лемешоу, по результатам применения которой принимается решение о необходимости выравнивания полученных результатов или ее отсутствии. Также выявляется присутствие / отсутствие образцов корреляции среди рассмотренных предикторов на основе функции *lrm*, что определяет меру надежности используемой модели и позволяет получить доверительные интервалы расчетных коэффициентов. Данный подход отражает новизну работы и позволяет выявить факторы, определяющие выбор того или иного понятия, исходя из объективных семантических критериев. Интеракции рассматриваются на 4 уровнях: научный, литературный, новостной и разговорный. Итоги работы позволяют дополнить содержание слов *list* и *catalog* и выявить их динамику.

Ключевые слова: корпусная лингвистика, логистическая регрессия, сравнительный анализ, семантика, синоним, *список*, *каталог*.