Five statistical models for Likert-type experimental data on acceptability judgments

Anna Endresen anna.endresen@uit.no & Laura A. Janda laura.janda@uit.no

CLEAR group
(Cognitive Linguistics: Empirical Approaches to Russian)
University of Tromsø: The Arctic University of Norway

Study of marginal change-of-state verbs in Russian (e.g. *ukonkretit'* 'concretize')[1]

Marginal (possible) word

- is attested at least once;
- is not established in standard language;
- is a spontaneous creation generated on the fly, on a certain occasion;
- is generated on the basis of a productive morphological pattern;
- is analyzable and semantically transparent.

Experimental design: score-assignment test

The task: Evaluate the marked word using one of the statements.

Давно пора как-то опричить наше общение более мягкими выражениями.
'It's high time we made our interaction respectable by using kinder statements.'

□ 5 points - Это совершенно нормальное слово русского языка.
'This is an absolutely normal Russian word.'
□4 points - Это слово нормальное, но его мало используют.
'This word is normal, but it is rarely used.'
□ 3 points - Это слово звучит странно, но, может быть, его кто-то использует.
'This word sounds strange, but someone might use it.'
□ 2 points - Это слово звучит странно, и его вряд ли кто-то использует.
'This word sounds strange and it is unlikely that anyone uses it.'
□ 1 point - Этого слова в русском языке нет.
'This word does not exist in the Russian language.'

Stimuli: 60

- 3 word types: 20 STANDARD verbs with high token frequency vs. 20 MARGINAL verbs with minimal token frequency vs. 20 NONCE verbs with zero attestations.

[1] For more details on the experimental design see Endresen 2014 (in Eng) and http://munin.uit.no/bitstream/handle/10037/5476/article.pdf?sequence=1 (in Rus). The data and R code for the five statistical models discussed in this talk are available at TROLLing (The Tromsø Repository of Language and Linguistics): http://hdl.handle.net/10037.1/10078

#	O- factitive	Gloss	Freq RNC 1950-2012	U- factitive	Gloss	Freq RNC 1950-2012
1	*ob''jasnit'*	clarify	18,149	*utočnit'*	define more precisely	2,860
2	*oblegčit'*	simplify, lighten	1,802	*umen'šit'*	reduce	2,010
3	*oslabit'*	weaken, loosen	1,401	*uskorit'*	speed up	2,008
4	*okruglit'*	express in round numbers	939	*ulučšit'*	improve	1,899
5	*obogatit'*	enrich	800	*uprostit'*	simplify	1,350
6	*ožestočit'*	harden, obdurate	686	*ukorotit'*	make shorter	787
7	*osložnit'*	complicate	410	*usložnit'*	complicate	311
8	*ogolit'*	denude	387	*uteplit'*	make warmer	205
9	*osčastlivit'*	make happy	343	*uplotnit'*	compress	201
10	*osvežit'*	freshen	280	*uxudšit'*	make worse	199

Table 1: Standard change-of-state verbs used in experiment (control group 1).

#	O- factitive	Gloss	Freq RNC 1950-2012	U- factitive	Gloss	Freq RNC 1950-2012
1	*omeždunarodit'*	internationalize	1	*uvkusnit'*	make tastier	1
2	*opoxabit'*	profane, pollute	1	*umedlit'*	make slower	1
3	*opriličit'*	make decent	1	*ukrasivit'*	make prettier	1
4	*oser'ëznit'*	make serious	1	*user'ëznit'*	make more serious	1
5	*ostekljanit'*	make glassy	1	*ukonkretit'*	make more concrete	1
6	*oržavit'*	corrode	2	*usovremenit'*	make more modern	1
7	*osurovit'*	make rigorous	2	*ustrožit'*	make stricter	3
8	*obytovit'*	vulgarize	3	*ucelomudrit'*	make more innocent	3
9	*ovnešnit'*	externalize	4	*uprozračit'*	make more transparent	4
10	*omuzykalit'*	musicalize	4	*udorožit'*	make more expensive	8

Table 2: Marginal change-of-state verbs (possible words) used in experiment (tested group).

#	**O- factitive**	**U- factitive**
1	*osurit'*	*usaglit'*
2	*otovit'*	*utulit'*
3	*oduktit'*	*udamlit'*
4	*ogabit'*	*uguzvit'*
5	*okočlit'*	*ukampit'*
6	*ošaklit'*	*ušadrit'*
7	*očavit'*	*učopit'*
8	*oblusit'*	*uloprit'*
9	*obnomit'*	*unokrit'*
10	*obmomlit'*	*umarvit'*

Table 3: Nonce change-of-state verbs used in experiment (control group 2).

Presentation of the stimuli:
In the experiment, all change-of-state verbs are presented as perfective infinitives in contexts:

- For standard and marginal verbs we are using real contexts from the Russian National Corpus (www.ruscorpora.ru), often shortened.
- The contexts of nonce verbs are based on corpus contexts of real verbs with meanings similar to those that are assumed for nonce verbs.

- All 60 change-of-state verbs used in the experiment are **deadjectival**.
- All standard and marginal change-of-state verbs chosen for experiment are **morphologically and semantically transparent** and analyzable and have a clear existing adjectival base.

3 research questions

PREDICTOR 1: PREFIX

- Does the more productive prefix O- form more acceptable novel marginal verbs than the less productive prefix U-?

PREDICTOR 2: AGE OF SPEAKER

- Does the speakers' leniency regarding marginal verbs correlate with age? Do adults (25-62 year old, N=51) have more conservative judgements than children (14-17 year old, N=70)?

PREDICTOR 3: WORD TYPE

- Are MARGINAL verbs of the two rival patterns (O- and U-) perceived more like STANDARD or more like NONCE verbs?

Dependent variable: a response score assigned to a stimulus.
Tested independent variables: Prefix, Age, Word type, Gender.

Central tendencies in data distribution

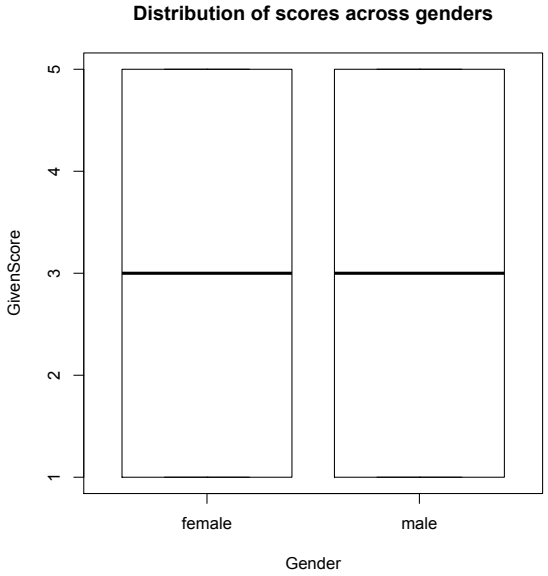


Figure 1: Impact of Gender.

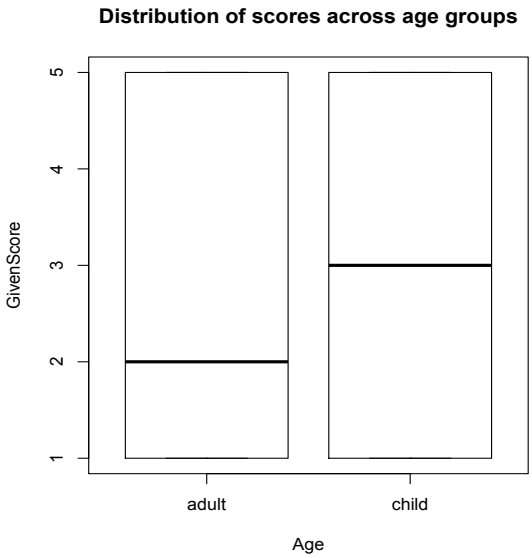


Figure 2: Impact of Age.

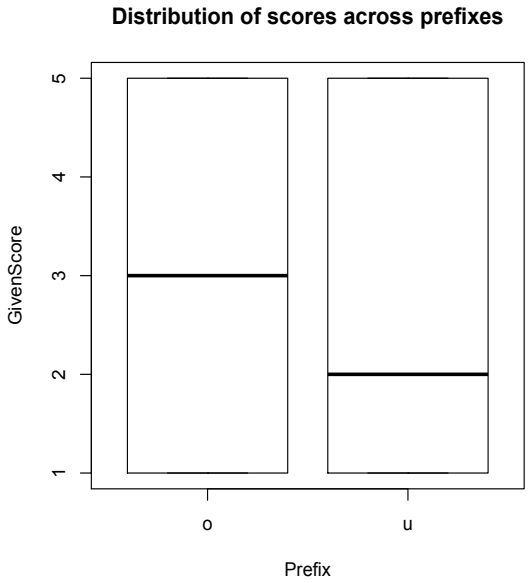


Figure 3: Impact of Prefix (O- vs. U-).

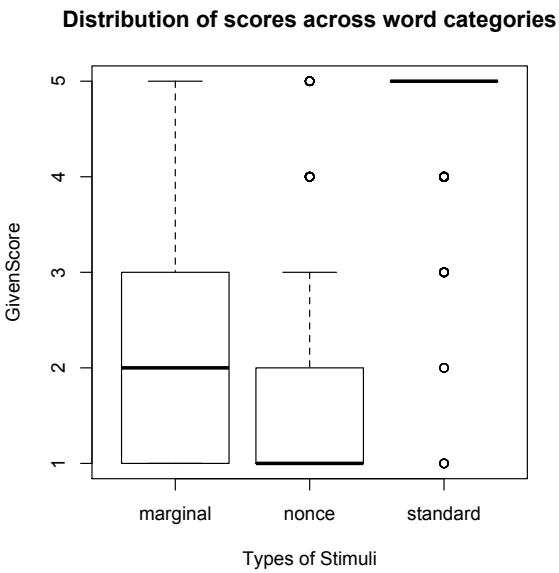


Figure 4: Impact of Word type.

Statistical modeling of experimental results

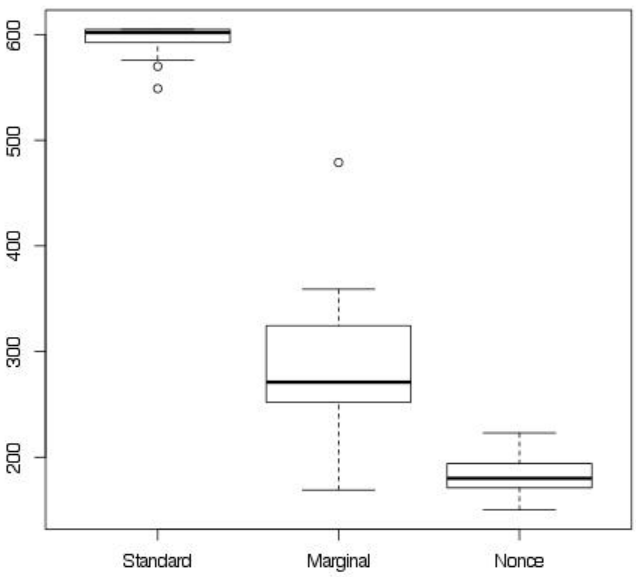
The null hypothesis: no statistically significant correlations among the variables.
The alternative hypothesis: significant correlations among the variables exist.

Type of test	Name	Type of data	Significant factors
Parametric	ANOVA	For interval data	WordType
	Ordinal logistic regression	For ordinal data	WordType >>> AgeGroup > Prefix
	Regression mixed-effects model	For ordinal data	WordType >>> AgeGroup
Non-parametric	Regression tree & Random forests	For numerical ordinal data	WordType >>> AgeGroup > Prefix
	Classification tree & Random forests	For categorical data	WordType >>> Prefix > AgeGroup

Table 4: Overview of five statistical models.

Model 1: ANOVA

- R script available at http://ansatte.uit.no/laura.janda/PossWords/PossWords.R
- ANOVA results overall: F= 546, df = 2, p-value < 2.2e-16 (the difference between distribution of acceptability scores across the three classes is significant). This shows that three categories of words (Standard vs. Marginal vs. Nonce) are perceived by speakers differently.



Standard Verbs
MAX = 605
MEAN = 595
MIN = 549
stand dev = 15
variance = 235

Marginal Verbs
MAX = 479
MEAN = 286.4
MIN = 169
stand dev = 67
variance = 4446

Nonce Verbs
MAX = 223
MEAN = 183.4
MIN = 150
stand dev = 19
variance = 360

Figure 5: Distribution of the three types of words in terms of acceptability ratings.

T-test RESULTS for **standard vs. marginal** words:
t = 20, df = 21, p-value = 3.173e-15, 95% confidence interval is 277 340
T-test RESULTS for **marginal vs. nonce** words:
t = 7, df = 22, p-value = 1.098e-06, 95% confidence interval is 71 135

- Marginal verbs are evaluated by speakers more like nonce verbs than standard verbs (This suggests that speakers are more sensitive to frequency than to semantic transparency).

Model 2: Ordinal Logistic Regression

- Logistic regression is a well established robust and powerful statistical technique that is widely used for multifactorial analysis (Strobl et al. 2009: 323; Baayen et al. 2013: 260).
- Baayen (2008: 208): a logistic regression analysis is appropriate for those dependent variables that are dichotomous, i.e. contain binomial values.
- In our case we are dealing with a multinomial dependent variable with five ordered values. For such ordered dependent variables it is appropriate to use the kind of logistic regression which is specifically designed for ordinal data analysis – an **Ordinal Logistic Regression** (Baayen 2008: 208-214).[2]
- We used the packages languageR, rms[3], and MASS and the function lrm(). The analysis was conducted using R version 2.15.0.

- In the Ordinal Logistic Regression analysis we approach the dependent variable Score as **ordinal data**. This analysis shows that three factors are **statistically significant predictors of acceptability scores – WordType and AgeGroup** (with p-values <0.0001, or ***) **and Prefix** (with p-value=0.0195, or *). The impact of Gender was found insignificant: Chi-Square= 0.33, df = 1, p-value = 0.56.
- The final and most optimal model included three factors that have significant effect on the choice of the Score – WordType, AgeGroup, and Prefix.

Factor	Chi-Square	Degrees of freedom	*p*-value
AgeGroup	59.28	1	< .0001
Prefix	5.45	1	0.0195
WordType	3415.95	2	< .0001
TOTAL	3425.06	4	< .0001

Table 5: Outcome of the Ordinal Logistic Regression: Wald Statistics.

- From comparison of the chi-square values we can conclude that **the impact of WordType accounts for most of data**, while **the other two factors are very minor.**

- The summary of the Logistic Regression Analysis provides the measures of predictive strength of the model. All three important measures – C[4], Somer's

[2] In order to make the outcome variable Score an ordered factor with levels 1<2<3<4<5 we used the function *ordered()*: dat$Score=ordered(dat$Score, levels=c("E","D","C","B","A")).
[3] Because the package 'Design' was removed from the CRAN repository, we used the package 'rms' instead.

Dxy[5], and the R^2 index (Harrel 2001: 248; Baayen 2008: 204) – are high and indicate **the high predictive strength of the model**:

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	7260	LR chi2	7618.29	R2	0.689	C	0.855
max |deriv|	7e-12	d.f.	4	g	3.136	Dxy	0.710
		Pr(> chi2)	<0.0001	gr	23.016	gamma	0.754
				gp	0.380	tau-a	0.518
				Brier	0.119		

Table 6: Outcome of the Ordinal Logistic Regression.

Model 3: Regression Mixed-Effects Model for Ordinal Data

- The model that can generalize over the bias of individual subjects and stimuli and determine a tendency which predominates over random effects.

Fixed-Effects factors:
WordType: standard, marginal, nonce
AgeGroup: child, adult
Prefix: O-, U-
Gender: male, female

Random-effects factors:
Subject: 121 persons
Stimulus: 60 verbs

High variation across subjects: contradictory acceptability judgments of marginal words:

Marginal factitive	Gloss	Number of subjects who gave				
		5 scores (normal word)	4 scores	3 scores	2 scores	1 score (does not exist)
usovremenit'	'modernize'	22	26	27	18	28
opriličit'	'make decent'	9	25	33	22	31

Table 7: Variation across subjects regarding the same marginal stimuli.

- Mixed-effects models are primarily used to explore data with nominal binomial dependent variables (0/1, A/B) (e.g. Tagliamonte & Baayen 2012) or continuous numerical dependent variables, for example reaction time (e.g. Baayen 2008: 242-302).
- In order to account for a multinomial ordinal dependent variable by means of a mixed-effects model, we used the package Ordinal in its latest version 2013.9-13[6] available in R version 3.0.2.

[4]C is the index of concordance between the predicted probability and the observed response. According to Baayen (2008: 204), "[w]hen C takes the value 0.5, the predictions are random, when it is 1, prediction is perfect. A value above 0.8 indicates that the model may have some real predictive capacity". In our case, C is higher than 0.8, which suggests that the model has a high predictivity.

[5] Somer's Dxy is an index of a rank correlation between predicted probabilities and observed responses. According to Baayen (2008: 204), "this measure <…> ranges between 0 (randomness) and 1 (perfect prediction)."

[6] See the description at http://cran.r-project.org/web/packages/ordinal/index.html

- We used the function clmm() which can handle the crossed random-effects structure of two factors – Subject and Stimulus.[7]
- Technically the Regression Mixed-Effects Model is a parametric model, but it does not assume a normal distribution for the response.

Groups	Name	Variance	Standard Deviation
SubjectCode	(Intercept)	1.091	1.045
Stimulus	(Intercept)	1.043	1.021

Table 8: Random-effects factors.

	Estimate	Std. error	z value	Pr(>|z|)	
AgeGroup-child	0.5803	0.2013	2.883	0.00394	**
WordType-nonce	-1.7791	0.3292	-5.405	6.48e-08	***
WordType-standard	7.4203	0.3712	19.991	< 2e-16	***

Table 9: Fixed-effects factors.

- The impact of **Gender** and **Prefix** is found insignificant in terms of predicting the dependent variable Score.
- After elimination of these factors, the most optimal fitted model indicated the **significant effects of two factors – WordType and AgeGroup**.
- The effect of WordType is more significant than that of AgeGroup.

Models 4 and 5: Regression and Classification Trees (CART) & Random Forests

- Classification and Regression Trees is a new method that is quickly gaining popularity in genetics, medicine (Strobl et al. 2009: 324), social sciences, and linguistics (Tagliamonte & Baayen 2012 and Baayen et al. 2013).
- Classification and Regression Trees is a non-parametric statistical technique which is appropriate for non-interval data.
- CART analysis provides a powerful tool to explore an ordinally scaled dependent variable (Faraway 2006: 253-268; Baayen 2008: 148-164).
- The Trees method has many advantages and has proven to give robust results, comparable with more traditional models like Logistic Regression, and even to give more accurate predictions, especially regarding complex multifactorial interaction effects (Baayen 2008: 154; Baayen et al. 2013).
- In a linear model like Logistic Regression the predictors are analyzed in a linear way in order to model their impact on the response (dependent) variable. By contrast, nonparametric regression models like Trees do not employ linearity and are often more flexible in modeling combinations of predictors (Faraway 2006: v).
- Trees do not hold any assumptions about the normal distribution of the response variable (as opposed to the logistic regression model) and can cope with any data structure and type and are highly recommended for unbalanced datasets.
- Variable importance ranking is available via the extension of CART to the Random Forest approach.
- Random Forest produces a variable importance scale to compare all tested predictors with each other in terms of their strength.

[7] We are indebted to Rune Haubo Bojesen Christensen for pointing out this possibility.

The outcome of the CART analysis is a graphically plotted "tree". It represents an algorithm of data partitioning which consists of recursive binary splits. The Tree outlines a decision procedure of predicting the values of the dependent variable:

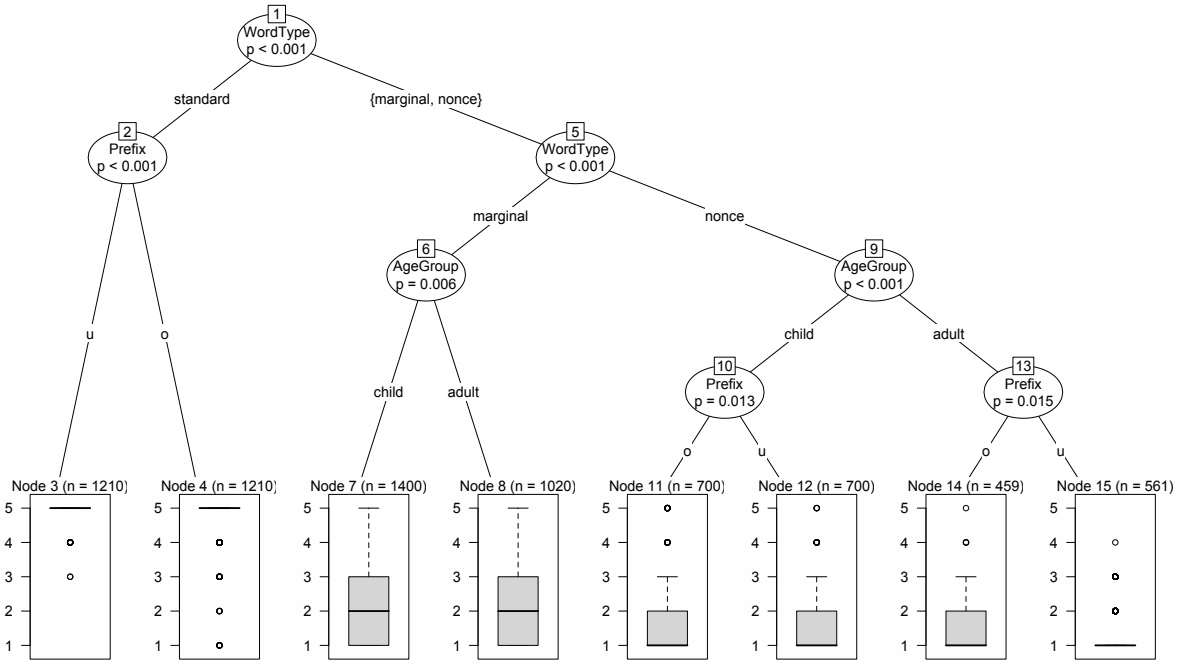


Figure 6: Regression tree of acceptability ratings: scores are treated as numerical ordinal data - from 5 points to 1 point.

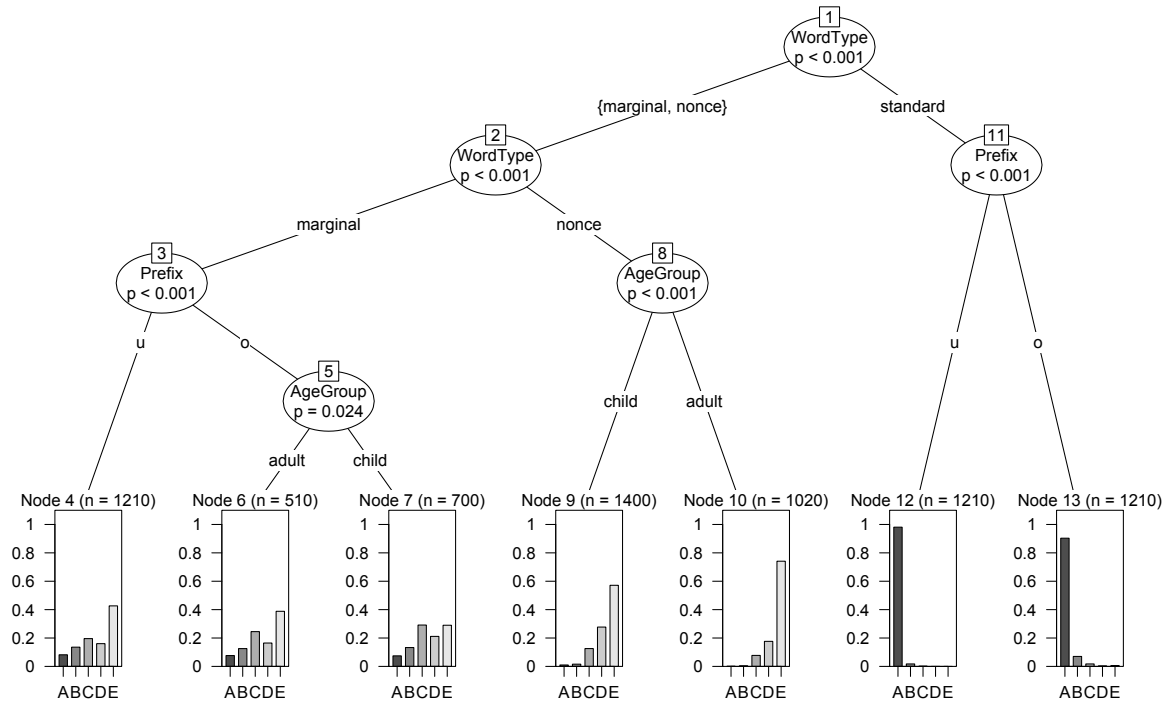


Figure 7: Classification tree of acceptability ratings: scores are treated as categorical data: A-score "5"; B-score "4"; C-score "3", D-score "2", E-score "1".

- Both trees show high-level interactions of WordType, AgeGroup, and Prefix.
- In both trees WordType is the most important factor, while Prefix and AgeGroup play their roles locally, making rather slight differences.
- The effects of AgeGroup and Prefix are statistically significant and optimal only within the scope of each local split.

- A Ctree treats the values of a dependent variable as a categorical scale.
- An Rtree applies to numerical dependent variables (Baayen 2008: 148).
- Because Ctree and Rtree handle different kinds of data, they differ in mechanisms of data partitioning:
    - Ctree makes splits according to the principle of increasing purity of a node: after each split the subgroups of data observations should become purer, or more of the same kind.
    - An Rtree employs the residual sum of squares as a criterion for splitting the nodes (Faraway 2006: 261). In addition, Rtree also computes the mean within each partition (ibid: 261).

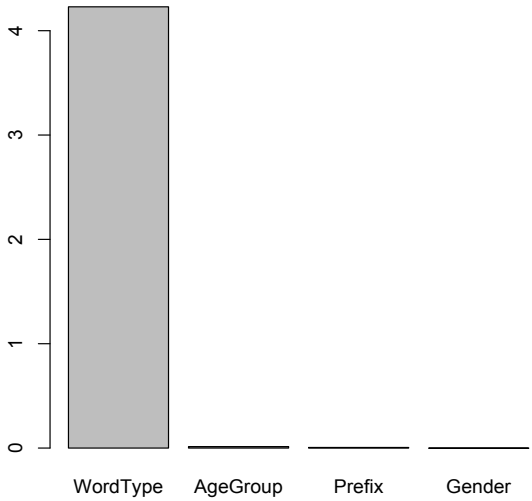


Figure 8: Variable importance scale for ordinal data (A>B>C>D>E).

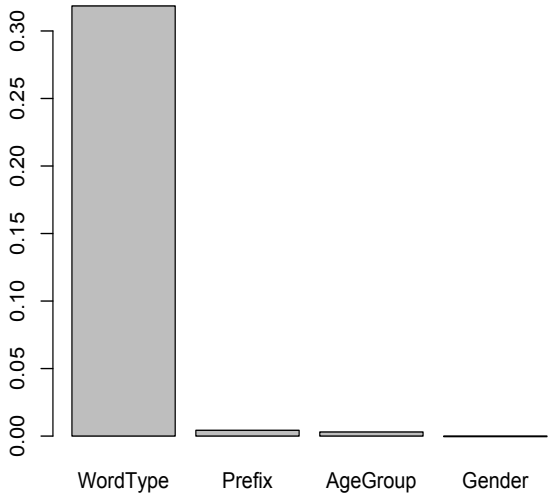


Figure 9: Variable importance scale for categorical data (A, B, C, D, E).

- Both plots depict the same four factors and arrange them almost identically.
- Both plots show that WordType is absolutely the strongest predictor, while the impact of other factors is close to zero.
- Both plots show that Gender is the weakest predictor of all (it appeared in neither of the Trees).
- Prefix and Age Group are ranked differently: Forest analysis of categorical data (Figure 9) suggests that Prefix is slightly stronger than AgeGroup, while Forest analysis of ordinal data (Figure 8) supports the reverse ranking, with a stronger impact of AgeGroup followed by Prefix.
- However, the difference between the importance scores of these two factors is very small in both plots.

Conclusions

- The experimental study targets those change-of-state verbs that have marginal status in Modern Russian.
- We tested whether the prefix (O- vs. U-), gender and age of speakers, and word type correlates with higher or lower acceptability of novel coinages in perception of native speakers.
- We approached the data from different perspectives, applying both parametric and non-parametric statistics.

- Parametric tests provide outcomes comparable with non-parametric models.
    - All models identify WordType as the major predictor.
    - The differences concern the factors AgeGroup and Prefix that have very small impact.
    - The five applied models focus on different aspects of data.

- We suggest that the non-parametric Classification Tree model is the most insightful and fruitful regarding this data.
    - This model is most informative regarding marginal verbs that are the focus of this study.
    - This model demonstrates that the importance of a factor can belong to different "levels": what is crucial at the level of a local split (AgeGroup and Prefix) might have very small overall predicting power considering the entire dataset, while other factors (like WordType) can determine the major trend of data distribution, as we saw in the major split of the Trees and the highest bar in the Random Forest plots.
    - The outcome of Random Forest analyses indicates that AgeGroup and Prefix do have some importance but their effect is very small. This effect is revealed in high level interactions of the factors.

- The major role of **WordType** is supported by Trees, Random Forests, ANOVA test, Ordinal Logistic Regression Model and Ordinal Mixed-Effects Regression Model.
- The relatively small importance of **Prefix** revealed by the Random Forest analysis is comparable with the outcome of Ordinal Logistic Regression, where Prefix is the least significant of three factors; and is also parallel to the result of Ordinal Mixed-Effects Regression, where Prefix is not found to be significant at all.
- The low predictive strength of **AgeGroup** revealed by Random Forest corresponds to what was found by ANOVA test. This contradicts with the result of the Ordinal Logistic Regression and the Mixed-Effects Regression analyses, where the effect of AgeGroup was found to be statistically significant.

- In terms of acceptability, **marginal words pattern closer to nonce words than to standard words.** This finding might be explained by the linguistic culture specific for Russia, which implies strong linguistic norms and in particular strong concern for the purity of proper literary language.

References

Baayen, R. Harald. 2008. *Analysing linguistic data. A practical introduction to Statistics using R*. Cambridge University Press.

Baayen et al. 2013 – Baayen, R. Harald, Laura A. Janda, Tore Nesset, Anna Endresen, Anastasia Makarova. 2013. Making choices in Slavic: Pros and cons of statistical methods for rival forms. In *Russian Linguistics 37* (Special issue "Space and Time in Russian Temporal Expressions"). 253-291.

Bard, Ellen Gurman, Dan Robertson and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. In *Language* 72,1. Pp. 32-68.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge University Press.
Bermel, Neil & Luděk Knittl. 2012. Corpus frequency and acceptability judgements: A study of morphosyntactic variants in Czech. In *Corpus Linguistics and Linguistic Theory*, 8(2), 241-275.
Cohen et al. 2000. Research Methods in Education. 5th ed. London: Routledge Falmer.
Dąbrowska, Ewa. 2010. "Naive vs. expert intuitions: An empirical study of acceptability judgments". *The Linguistic Review* 27, 1-23.
Endresen, Anna. 2014. *Non-Standard Allomorphy in Russian Prefixes: Corpus, Experimental, and Statistical Exploration*. Doctoral dissertation. University of Tromsø: The Arctic University of Norway. Available at http://hdl.handle.net/10037/7098
Faraway, Julian J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science.
Jamieson, S. 2004. Likert scales: How to (ab)use them. In Medical Education. 38. 1212–1218.
Lavrakas, Paul J. 2008. Encyclopedia of Survey Research Methods. SAGE Publications.
Likert, Rensis. 1932. A Technique for the Measurement of Attitudes. PhD dissertation. Columbia University. NY: The Science Press. Published in series "*Archives of Psychology*", 140, 1–55.
Schütze, Carson T. (1996) *The Empirical Base of Linguistics: Grammaticality judgements and linguistic methodology*. Chicago and London. The University of Chicago Press.
Sorace, Antonella and Frank Keller. 2005. Gradience in Linguistic Data. In *Lingua* 115, 1497-1524.
Strobl et al. 2009. An introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. In Psychological Methods. 14.4. 323-348.
Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was*/*were* variation as a case study for statistical practice. In *Language Variation and Change*. 24.2. 135-178.

Appendix: sample of stimuli

Standard stimuli

(1) *S pomošč'ju ètoj očiščajuščej maski možno legko uvlažnit' kožu i* ***osvežit'*** *cvet lica.*
'By means of this cleansing mask one can easily moisturize the skin and **freshen** the complexion.'

(2) *Novaja sistema pozvoljaet* ***uskorit'*** *dostavku gruzov i povysit' bezopasnost' personala.*
'The new system makes it possible to **speed up** transportation and to increase staff safety.'

Marginal stimuli

(3) *Gollivud uxitrilsja* ***opoxabit'*** *počti vse šedevry literatury.*
'Hollywood has managed to **profane** almost all masterpieces of fiction.'

(4) ***Usovremenit'*** *arxitekturu v gorodax Rossii možno bylo by putem snosa vetxix domov v centre mnogix gorodov.*
'It could be possible to **modernize** the architecture in Russian cities by demolishing shabby houses in many city centers.'

Nonce stimuli

(5) *Novye komp'juternye igry mogut* ***otovit'*** *ljubogo: na èto rabotaet i grafika, i cvetovaja gamma, i sjužet.*
'New computer games can **affect** anyone: for this purpose they employ a certain graphic design, color range, and plot.'

(6) *Esli vy voz'mete s soboj sobaku, pridetsja* ***ukampit'*** *stoimost' èkskursii na 40 rublej.*
'If you take the dog along, we will have to **change** the price of the tour by forty rubles.'