Journal of Research Design and Statistics in Linguistics and Communication Science

JRDS (PRINT) ISSN 2052-417X
JRDS (ONLINE) ISSN 2052-4188

Article

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

Anna Endresen and Laura A. Janda

Abstract

This paper contributes to the ongoing debate over Likert scale experiments, in particular the issues of how to treat acceptability judgment data (as ordinal or interval) and what statistical model is appropriate to apply. We analyze empirical data on native speakers' intuitions regarding marginal change-of-state verbs in Russian (e.g. ukonkretit' 'concretize', ovnešnit' 'externalize') and compare the outcomes of five statistical models (parametric and non-parametric tests): (1) ANOVA; (2) Ordinal Logistic Regression Model; (3) Mixed-Effects Regression Model for Ordinal data; (4) Regression Tree and Random Forests Model; and (5) Classification Tree and Random Forests Model. We make four claims: (1) all five models are appropriate for this data to a greater or lesser degree; (2) overall, the outcomes of parametric and non-parametric tests applied to this data provide comparable results; (3) Classification Tree and Random Forests Model is the most appropriate, informative, and user-friendly regarding this data; and (4) the use of a culturally entrenched grading scale is an advantage.

Keywords: Likert scale; acceptability judgment; experiment; marginal verb; prefix; Russian

Affiliation

UiT The Arctic University of Norway, Postboks 6050 Langnes, 9037 Tromsø.
email: anna.endresen@gmail.com (corresponding author)

JRDS VOL 3.2 2016 217–250 https://doi.org/10.1558/jrds.30822

EQUINOX
EQUINOXONLINE

1. Introduction

Likert-type scales are widely used in linguistic experiments as a technique for elicitation of acceptability scores. In such studies, subjects are presented with a ranked set of points (usually five or seven) where the top and the bottom ends are descriptively categorized, for example, as 'perfectly normal' and 'unacceptable'. There is a controversy in the literature as to whether one can assume equal intervals between values on such a scale of acceptability judgments, and therefore treat this data as interval and apply parametric statistics like ANOVA and Logistic Regression. Although parametric tests have commonly been applied to Likert-derived data (Lavrakas, 2008; Strobl *et al.*, 2009; Dąbrowska, 2010; Bermel and Knittl, 2012), some scholars find this practice illegitimate and erroneous in terms of data analysis and interpretation (Jamieson, 2004; Grilli and Rampichini, 2012). In this article we address the debate about appropriate statistical models for Likert-type data by comparing the outcomes of parametric and non-parametric statistical models applied to the same data set.

We used a Likert-type scale in order to compare native speakers' intuitions regarding 60 standard (common), marginal (rare), and nonce (non-extant) verbs in Contemporary Standard Russian. We report on an experimental study that recruited 121 participants and tested whether acceptability scores correlate with four predictor variables: (1) the prefix of the verb (more productive O- vs. less productive U-); (2) speaker's age (middle school children vs. adults); (3) gender (male vs. female); and (4) word type (standard verbs with high token corpus frequency, marginal verbs with minimal token corpus frequency, and nonce verbs with no corpus attestations).

We suggest that these data present an ideal ground to test the power of competing statistical models. In doing so we attempt to answer two questions:

1. Are the outcomes of parametric and non-parametric statistical tests comparable?
2. Which model is the most appropriate, informative, and user-friendly given the data we collected?

This study has the potential to provide useful insights on the use of Likert and Likert-type scales. The ambition of this article goes beyond a specific data set and acknowledges a range of possibilities for statistical analysis.

In this article we discuss in detail the five models listed in Table 1. The leftmost column (1) divides the five models into parametric vs. non-parametric tests, column (2) names the models, column (3) specifies what type of data measurement is appropriate for each model, and the rightmost column spells out the outcome of each model applied to our data in terms of the significant factors ordered according to their relative importance.

Table 1: Overview of five statistical models applied to our data set[1]

Type of test	Name of the model	Type of data measurement of the response variable	Outcome: Significant factors
Parametric	ANOVA	Interval data	WordType
	Ordinal logistic regression	Ordinal data	WordType >>> AgeGroup > Prefix
	Mixed-effects Regression model	Ordinal data	WordType >>> AgeGroup
Non-parametric	Regression tree and Random forests	Numerical ordinal data	WordType >>> AgeGroup > Prefix
	Classification tree and Random forests	Categorical data	WordType >>> Prefix > AgeGroup

As shown in Table 1, we apply three parametric models including ANOVA, Ordinal Logistic Regression, and Mixed-Effects Regression. ANOVA treats our data set as interval data, whereas Ordinal Logistic Regression and Mixed-Effects Regression models are specifically designed to handle ordinal data. Note that Mixed-Effects Regression is a nonlinear model in its parameters (cf. Christensen and Brockhoff, 2013: 59) as opposed to Ordinal Logistic Regression. We compare the outcomes of three parametric models with two non-parametric models: Regression Tree designed for numerical ordinal data and Classification Tree designed for categorical data.

All five statistical models have the power to handle multifactorial analysis, but they make different assumptions about the data they are applied to. Based on the study we conducted, we make four claims. First, we propose that each of these models is appropriate for Likert-type data to a greater or lesser degree depending on how seriously one takes the 'intervalness' assumption (see section 2). Second, we find that overall, the outcomes of parametric and non-parametric tests applied to this data set provide comparable results (see section 4 for details). Third, we advocate the latter model (Classification Tree and Random Forests) as the most fruitful, appropriate, informative, and user-friendly regarding the data we collected (see section 5). This model makes the least assumptions about Likert-type data and at the same time provides the most informative insights about the focus of this study, that is the perception of marginal verbs. In particular, whereas each model identifies WordType as the major predictor, Classification Tree additionally

shows that AgeGroup and Prefix have significant effects within local subsets of data, where the factors interact. Fourth, we suggest that the use of a culturally entrenched grading scale in Likert-type experiments is an advantage, because it has a normative effect on the interpretation of scale points (section 3). We propose that an entrenched grading scale gives better control over subjects' intuitions and shields the results from unwanted additional opaque variables.

The article is organized as follows. In section 2 we discuss the use of Likert-type scales and address the debate regarding what kinds of statistical tests are appropriate for this kind of data. In 3, we present our study of native speakers' intuitions regarding marginal verbs in Contemporary Standard Russian. In 4, we subject the collected data to five statistical models and compare their outcomes in section 5. The contribution of this article is summarized in section 6.

We share all supplementary materials for this article at TROLLing, i.e. the Tromsø Repository of Language and Linguistics (http://opendata.uit.no/). The experimental questionnaire, the database of collected responses, and R code for the five statistical models discussed in this article can be freely accessed at https://opendata.uit.no/dataset.xhtml?persistentId=hdl:10037.1/10256.

2. Preliminaries: The controversy over Likert and Likert-type scale data

The Likert scale is named after its inventor, American psychologist Rensis Likert, who proposed a method for measuring individual attitudes by collecting people's responses in terms of how much they agree to a given statement (Likert, 1932). Thus, strictly speaking, a typical Likert scale is a scale of agreement, as in (1), where the choices form a continuum from 'strongly disagree' to 'strongly agree':

(1) *The instruction of this university course was easy to follow.*

strongly disagree	disagree	undecided	agree	strongly agree
◯	◯	◯	◯	◯
1	2	3	4	5

Lavrakas (2008: 429) distinguishes the *Likert scale proper* (1) from similar scales termed *Likert-like* or *Likert-type scales* that allow respondents to indicate the degree of importance, frequency, quality, or satisfaction, as in (2):

(2) *How satisfied are you with the security control at the airport?*

Very dissatisfied	Somewhat dissatisfied	Neither dissatisfied nor satisfied	Somewhat satisfied	Very satisfied
◯	◯	◯	◯	◯
1	2	3	4	5

However, despite this terminological distinction, both Likert and Likert-type scales belong to the same family of methods that 'ascribe quantitative values to qualitative data' in order to make it amenable to statistical analysis (Dubois, 2013: 132).

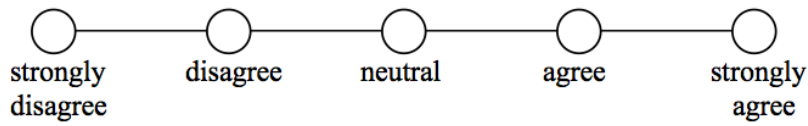
There is a long-term debate regarding the level of measurement of Likert-derived data, in particular whether such data constitute an ordinal or an interval scale. This matters because the statistical techniques used for interval variables are not appropriate for ordinal variables. In particular, interval data can be subjected to parametric tests (like calculation of mean and variance), while ordinal data can only be explored via non-parametric tests like the chi-squared test (Cohen *et al*., 2000: 317; Cantos Gómez, 2013: 236). The use of the wrong statistical test arguably leads to incorrect conclusions about data.

Strictly speaking, the intervals between values on a Likert scale are not necessarily equal, but many researchers assume that they are. Cohen *et al.* (2000: 317) and Jamieson (2004) object against assuming an interval scale for Likert-type categories. They find it illegitimate to use parametric statistics for data obtained via Likert scales. Yet, Jamieson observes that in medical and social sciences it has become 'a common practice to assume that Likert-type categories constitute interval-level measurements' (Jamieson, 2004: 1212). Similarly, Strobl, Malley, and Tutz (2009: 323) mention the fact that 'ordinally scaled variables, which are particularly common in psychological applications, are often treated as if they were measured on an interval or ratio scale'. Lavrakas (2008) states along the same lines that while 'it is common to treat Likert scales as interval level data, it is more conservative to view such data as ordinal'. Ordinal-level variables are generally considered challenging for statistics. The ordinal/interval scale-and-statistics controversy is a long-standing and continuing debate. Knapp (1990: 121) points out that the distinction between ordinal and interval scales of data measurement is often a challenge when one has to categorize a specific data set. Moreover, Knapp (1990: 121) suggests that a particular scale can be 'ordinal, less than ordinal, or more than ordinal', and that there are no agreed-upon rules for determining this.[2]

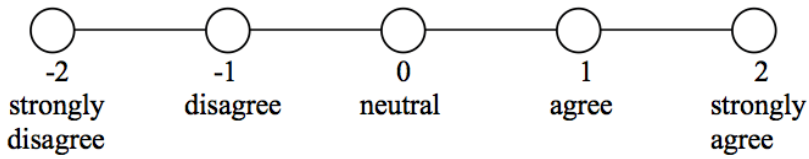
Technically, the response format where each item of the scale is a descriptive statement is at the ordinal or even categorical level of measurement. How-

ever, the key assumption behind using Likert-type scales is that the distances between each two adjacent items are of equal magnitude. The sense of equidistance is often reinforced graphically, as in (3), or by adding a set of numbers, as in (2) and (4). The ambition of these designs is to justify the use of the Likert scale as an interval-level measurement, in turn facilitating parametric statistics.

(3) *The instruction of this university course was easy to follow.*

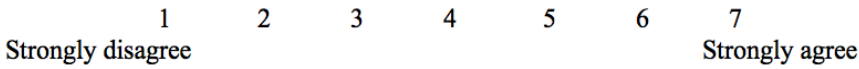


(4) *The instruction of this university course was easy to follow.*



Another solution that reinforces equal distances between points on a scale is a format where only endpoints of the scale are descriptively categorized and midpoints are not labeled, as in (5):

(5) *The instruction of this university course was easy to follow.*



This format was employed by Dąbrowska (2010) and Bermel and Knittl (2012) to elicit acceptability judgments, scaling evaluation of a linguistic form from *Unacceptable* to *Perfectly Normal*, as in (6):

(6) *<Linguistic form>*

Dąbrowska (2010: 8) points out that a number of studies (Jaccard and Wan, 1996; Labovitz, 1967; Kim, 1975) have argued that 'parametric tests are quite robust, so that violations of the intervalness assumption have relatively little impact on the results of the test'. Dąbrowska (2010: 8) states that 'the use of

parametric tests with data obtained using Likert scales has now become standard' (cf. similar observations in Blaikie, 2003; Pell, 2005). Dąbrowska (2010) herself uses a five-point Likert-type scale in elicitation experiments and analyzes the responses with ANOVA and *t*-tests. Similarly, Bermel and Knittl (2012) conduct an experiment using a seven-point Likert-type scale and explore their results with ANOVA statistics. However, both Dąbrowska (2010) and Bermel and Knittl (2012) additionally conduct non-parametric statistical tests to verify their results.

Rietveld and van Hout (2005: 136) claim that 'one has not to worry too much about the scale level of the data which is submitted to analysis of variance. <…> F ratios are not so much affected by violations of the assumptions of the interval measurement level as one often thinks.' Rietveld and van Hout conclude that 'analysis of variance can be applied to the data which are not strictly of the interval level'. Labovitz (1970: 515) ensures that 'Empirical evidence supports the treatment of ordinal variables as if they conform to interval variables <…>. Although some small error may accompany the treatments of ordinal variables as interval, this is offset by the use of more powerful, more sensitive, better developed, and more clearly interpretable statistics with a known sampling error.'

Christensen and Brockhoff (2013: 59) comment that normal linear models (regression and ANOVA) applied to ratings data treat inherently categorical data as continuous. Christensen and Brockhoff state that 'It is hard to quantify how this affects accuracy and consistency of parameter estimates as well as testing accuracy and power'. According to Christensen and Brockhoff, using parametric linear models for ordered categorical data 'can be a useful approximation if there are sufficiently many categories and not too many observations in the end categories'. In particular, Christensen and Brockhoff argue that linear models are inappropriate for scales with a small number of categories. As an alternative, Christensen and Brockhoff propose cumulative link models that treat ordinal data appropriately (see section 4.3).

Grilli and Rampichini (2012: 2) observe that 'in the social sciences the use of scoring systems to convert categories into numbers is common practice' for the reason that 'the statistical methods for quantitative variables are more powerful and easier to implement and interpret'. Grilli and Rampichini refer to a number of studies showing that the bias of analyzing ordinal data with methods for continuous data depends on the number of points on the scale and the skewness of the distribution. In this regard, Grilli and Rampichini claim that 'five is usually minimum to get an acceptable bias' and that 'the bias increases with the degree of skewness and may become large in the case of floor or ceiling effects, namely when the largest frequency corresponds to a category at the extremes of the scale'. Overall, Grilli and Rampichini advocate

'proper methods' for ordinal data variables, in particular various multilevel models designed to handle ordinal data (see section 4.3).

We suggest that the choice of a particular Likert-type scale for an experiment should be driven by the purposes of the study rather than by considerations of what type of scale is more likely to produce data that meet the assumptions of parametric tests. With the variety of statistical models available today, linguists can pursue their goals even if their data cannot strictly be interpreted as interval. The purpose of our study is to present the community of linguists with a number of options for a statistical analysis of a single data set. In this article we describe an experimental design where each point of the scale is both enumerated and descriptively labelled. This design best serves the purposes of our study. We describe our design in the next section and show that it is open to a number of possible interpretations in terms of the level of data measurement: categorical, ordinal, interval, and possibly even approaching ratio scale (with an interpretable zero point). We argue that our study thus represents an excellent ground for testing the applicability of various statistical models that include both parametric and non-parametric statistical tests.

3. Method: Elicitation of speakers' judgments of marginal verbs

The experiment targeted marginal new coinages attested in the Russian National Corpus (www.ruscorpora.ru). We focused on the productive derivation of verbs with the meaning 'make X be Y', where Y is an adjective that serves as the source for the derivation of a deadjectival verb denoting a change of state. For example, from the adjective *muzykal'nyj* 'musical' one can derive the verb *omuzykalit'* 'musicalize' denoting a change of state from non-musical to musical, and similarly from the adjective *konkretnyj* 'concrete' one can derive the verb *ukonkretit'* 'concretize' denoting the change of state from non-concrete to concrete. These marginal verbs are spontaneously produced by speakers and are not acknowledged in dictionaries. For the experimental study we chose verbs that are formed by the prefixes O- and U-, which are the two most productive prefixes used in this derivational pattern in Modern Russian (Townsend, 1968: 143; Endresen, 2014: 269).

3.1. Three research questions

The ultimate goal of the experiment was to test whether the acceptability scores assigned by speakers to verbal stimuli correlate with any of three factor variables: Prefix, AgeGroup, and WordType. In particular, we focused on the three questions listed in Table 2:

Table 2: Three research questions

Factor variable	Research question
1. Prefix	Does the prefix O- (which is the most productive prefix in this pattern) form more acceptable marginal verbs than the prefix U- (which is also very productive in this pattern, but less so than O-)?
2. AgeGroup	Does the speakers' leniency regarding marginal verbs correlate with age? Do adults (ages 25–62, *N* = 51) have more conservative judgments than children (ages 14–17, *N* = 70)?
3. WordType	Are MARGINAL verbs of the two rival patterns (O- and U-) perceived to be more like STANDARD or more like NONCE verbs?

First, we test whether the productivity of the prefix makes a difference in how marginal words are perceived by native speakers. We hypothesize that the two rival derivational patterns (prefixes O- vs. U-) are significantly different with regard to their relative naturalness to Russian speakers. In a previous corpus study, Endresen (2014: 269–284) showed that marginal change-of-state verbs prefixed in O- (e.g. *opoxabit'* 'profane') have two times higher type frequency than those prefixed in U- (e.g. *usovremenit'* 'modernize'). This suggests that in Modern Russian the O-pattern is more productive than the U-pattern and can be considered the default. We expect that novel marginal verbs formed by O- should be judged as more natural and acceptable than marginal verbs in U-. In other words, do speakers assign higher acceptability scores to marginal derivatives in O- and lower scores to derivatives in U-?

The second question that we addressed in the experiment is whether the speakers' leniency regarding marginal verbs changes with age. In particular, we were interested in two age groups of speakers – school age children and adults. Teenagers might be more liberal and open to unfamiliar words than adults, whose linguistic standards and preferences have already stabilized. In this regard, adults are expected to give more conservative judgments and be generally less willing to accept marginal words. We take the age of 25 as an approximate threshold for adulthood, because by this age most adults in Russia complete their education, enter the job market, and also outgrow colloquialisms typical for youth. On the other hand, we are particularly interested in 14–17-year-olds who are usually at the peak of implementing youth slang and are arguably very open to linguistic innovations.

Third, we compare *marginal words* to the two extremes – *standard words* that are well attested, conventionalized, and familiar to a language community, and *nonce words* that conform to the phonological rules but are not associated with any meaning. By definition, marginal words are not established in the standard lexicon. Rather, such words are spontaneous creations generated on the fly on a certain occasion. Marginal words are attested at least once in a corpus or elsewhere. Such words fill the gap between the actual and the impos-

sible in a language. On the one hand, marginal words exist because they are attested – they have been generated and recorded. On the other hand, marginal words do not exist, because most speakers have never heard them. In this study we want to find out whether speakers evaluate marginal lexemes as words or as non-words of their language.

3.2. Design: Benefits of a culturally embedded twofold scale

The experiment was designed as a score-assignment test. Each subject was presented with a total of 60 sentences and a rating system. Each sentence contained an underlined change-of-state verb, as illustrated in (7):

(7) *Davno pora kak-to opriličit' naše obščenie bolee mjagkimi vyraženijami.*
'It's high time we made our interaction respectable by using gentler expressions.'

The task was to evaluate the marked verb in a sentence according to a scale of acceptability judgments.[3] We used a numeric scale of five points combined with a categorical scale of evaluative statements shown below:

☐ 5 points – *Èto soveršenno normal'noe slovo russkogo jazyka.*
'This is an absolutely normal Russian word.'

☐ 4 points – *Èto slovo normal'noe, no ego malo ispol'zujut.*
'This word is normal, but it is rarely used.'

☐ 3 points – *Èto slovo zvučit stranno, no, možet byt', ego kto-to ispol'zuet.*
'This word sounds strange, but someone might use it.'

☐ 2 points – *Èto slovo zvučit stranno, i ego vrjad li kto-to ispol'zuet.*
'This word sounds strange and it is unlikely that anyone uses it.'

☐ 1 point – *Ètogo slova v russkom jazyke net.*
'This word does not exist in the Russian language.'

Thus, in our rating system we combined two types of scales: descriptive evaluative judgments provided a qualitative (categorical) scale, while a set of parallel scores from one to five formed a quantitative scale. Therefore, the subjects had to choose a combination of a statement and a score which described best their intuition about an underlined verb.

We suggest that the combination of numeric scores and descriptive judgments makes it possible to have better control over subjects' intuitions. Otherwise, subjects would have to improvise their interpretation of the five numerical scores. By contrast, the forced-choice system that we employ provides a uniform set of descriptions that each subject can rely on. In our exper-

iment this methodology helped to shield results from unwanted additional opaque variables and to collect more robust data.

The wording was meant to invite subjects to think generally, having in mind the whole language community. The evaluative statements are formulated in such a way that they maximally correspond to an ordinal scale with approximately comparable intervals between each two statements.

A rating scale subdivided into five points is culturally entrenched in Russia: this is the most common grading scale used in schools and universities. The scale consists of five grades, where grade '1' corresponds to the worst performance, while the top grade '5' corresponds to the best performance. We reflect the same gradation in our experimental scale, where the score '1' should be assigned to a word that does not exist in Russian, and the highest score of five points '5' should be assigned to a perfectly normal Russian word. This explains why we preferred to use the five-point scale instead of a three-point (Collins *et al.*, 2009) or seven-point scale (Bermel and Knittl, 2012). Moreover, the school evaluation system motivated us to use the scale of 1 to 5 instead of other options like −2, −1, 0, +1, +2.

We employed a vertical scale descending from 5 to 1 because this format corresponds to the iconic gradation 'the higher (spatially) – the better'.

3.3. Stimuli

Each questionnaire exposed subjects to three groups of stimuli – standard, marginal, and nonce change-of-state verbs prefixed in O- and U-. In order to limit the questionnaire to a manageable size, we used 20 stimuli in each group, with ten verbs prefixed in O-, and the other ten verbs in U-. An equal number of stimuli in each group was meant to counterbalance the group of marginal verbs and prevent the subjects from getting into a yea-saying or nay-saying mode (Schütze, 1996: 184). The prefix and the word-type conditions yield a total of 60 stimuli, where standard and nonce verbs were two groups of controls and distractors, whereas the 20 marginal verbs were the tested experimental items.

For this study we chose those marginal change-of-state verbs that were, like standard verbs, morphologically transparent and semantically analyzable. The marginal verbs differed from the standard verbs only in that they were not conventionalized and therefore mostly unfamiliar to an average speaker. On the other hand, being unfamiliar was a trait that marginal verbs shared with nonce verbs at the other extreme of the scale. However, marginal verbs were semantically felicitous, derived from a familiar adjective by means of a common word-formation pattern, while nonce verbs, by contrast, could not be associated with any existing adjectives.

In order to exclude other possible variables from the experimental conditions, all standard and marginal change-of-state verbs chosen for the exper-

iment had a clear adjectival base. None of the verbs had a parallel simplex verbal base: e.g. *ob"jasnit'* 'clarify' < *jasnyj* 'clear', but there is no **jasnit'* 'make clear'.[4] Nonce verbs followed the same morphological pattern of change-of-state verbs: they contained the same prefixes O- and U- and the verbalizing suffix *-i-*, but no recognizable root (see Table 5 for details).

The mode of presentation of all three types of stimuli was made uniform in terms of context. All stimuli were presented as perfective infinitives in a sentence which was borrowed or based upon a real sentence attested in the Russian National Corpus (www.ruscorpora.ru, henceforth RNC). We made sure that the contexts chosen for standard and marginal verbs were typical, neutral in register, and maximally supported the change-of-state meaning of the verb. The contexts for standard and marginal verbs were directly extracted from the corpus and were shortened in some cases. In a few cases a better context for a marginal verb was found via the search engines www.yandex.ru and www.google.ru. The contexts of nonce verbs were created to parallel the contexts of the standard and marginal verbs. All contexts of verbal stimuli used in the experiment are available at https://opendata.uit.no/dataset.xhtml?persistentId=hdl:10037.1/10256, where they are supplied with English translations.

Table 3 lists the standard verbs used in the experiment presented here in descending order of their token frequencies in the RNC.

Table 3: Standard change-of-state verbs used in experiment (control group 1)

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

All verbs in Table 3 have high token frequencies in the corpus. These frequencies are overall numbers of attestations of these verbs found in the Modern Subcorpus of the RNC, which includes the texts created in 1950–2012.

Table 4 provides a list of all marginal verbal stimuli employed in the experiment. When choosing these verbs we used two criteria – minimal token fre-

quency in the corpus and transparency of the word's derivational structure, in particular a clear semantic and structural association link with a base that speakers can easily rely on. In Table 4 the verbs are listed in increasing order of token frequencies – from one to eight corpus attestations.

Table 4: Marginal change-of-state verbs used in experiment (tested group)

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

Table 5 lists the nonce verbs used in the experiment. These verbs were adopted from the psycholinguistic experiments described in Endresen, 2014: Ch5.; Baayen, Janda, Nesset, Endresen and Makarova, 2013; Endresen, 2013.

Table 5: Nonce change-of-state verbs used in experiment (control group 2)

O-verb	U-verb	O-verb	U-verb	O-verb	U-verb
osurit'	*usaglit'*	*okočlit'*	*ukampit'*	*obnomit'*	*unokrit'*
otovit'	*utulit'*	*ošaklit'*	*ušadrit'*	*obmomlit'*	*umarvit'*
oduktit'	*udamlit'*	*očavit'*	*učopit'*		
ogabit'	*uguzvit'*	*oblusit'*	*uloprit'*		

The nonce verbs were created manually. They satisfy well-formedness constraints of Russian phonotactics and sound native-like to an average speaker. Table 5 demonstrates that each nonce verb in O- had a parallel nonce verb in U- which contains the same consonant of the base (*s, t, d, g*, etc.) but the base itself is not identical: e.g. *osurit'* and *usaglit', otovit'* and *utulit'*, etc. This was done in order to balance the set of nonce stimuli.

The stimuli were presented in a semi-random order that was the same for all participants. The first two warm-up sentences contained standard verbs,

while the third sentence introduced a marginal verb. We made sure that in each questionnaire there was no sequence of more than two adjacent sentences that introduced the same prefix or the same type of stimulus. This was done in order to prevent subjects from developing a uniform strategy that could bias their judgments.

3.4. Administration

The experiment was administered as a questionnaire with no time limits. The average time for completion of the questionnaire was 20 minutes. For children administration consisted of filling out a paper questionnaire form and was conducted in a school setting. Adults completed the survey over the internet, where they had to fill out a virtual questionnaire created in the software package http://www.questionpro.com. The use of an online questionnare form easily shared via internet is a common practice used in many recent surveys of acceptability judgments (Keller and Asudeh, 2001; Collins *et al*., 2009). The software made it possible to make sure that people who participated online took the survey only once.

The introduction to the experiment collected sociolinguistic information about subjects' gender, age, level of education, area of expertise, and place of residence. This part was followed by instructions about the task, the list of scores and statements, and an illustrative example with a standard change-of-state verb. The next section told the subjects that they would be exposed to both existing and non-existing words, that they would have to evaluate 60 words, and that the tasks contain no typos. This part informed the subjects that they should not worry about incorrect responses, because the task is not about spelling competence but rather about speakers' linguistic intuition. For all subjects, the scale of five scores accompanied with statements was given after each sentence.

3.5. Subjects

We recruited 121 subjects including 70 children and 51 adults. Among them there are 47 males and 74 females. All subjects are native speakers of Russian who grew up, received their education, and currently live in Russia.

4. Results: Statistical modeling of experimental results

4.0. Overview: Central tendencies of data distribution

Figures 1–4 plot the distribution of the dependent variable (acceptability scores assigned to stimuli) across the four tested independent variables – Prefix, AgeGroup of subjects, Gender, and WordType category. We include Gender here for the sake of comparison.

In each plot, the data is visualized in the shape of a rectangle, where the thick line indicates the median score. The box-and-whiskers plots neatly visu-

alize the central tendencies of the data distributions. Comparing these plots, we observe differences in the overall impact of the four factors.

Verbs prefixed in O- overall tend to receive higher acceptability scores compared to U-verbs (Figure 1): half of O-verbs received scores higher than '3', while half of U-verbs received scores higher than '2'. Children assign higher acceptability ratings than adults (Figure 2). Gender does not make any difference (Figure 3). Word types have three distinct patterns (Figure 4). Overall, marginal verbs received surprisingly low acceptability scores: half of the marginal verbs received the lowest scores of 1 and 2.

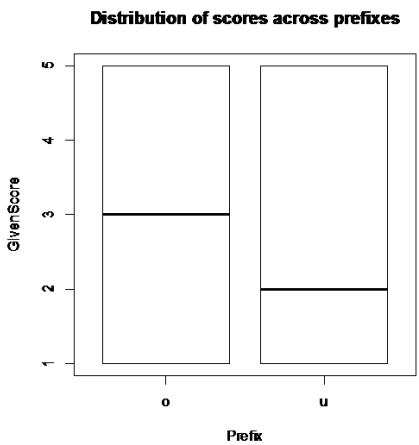


Figure 1: Impact of Prefix (O- vs. U-)

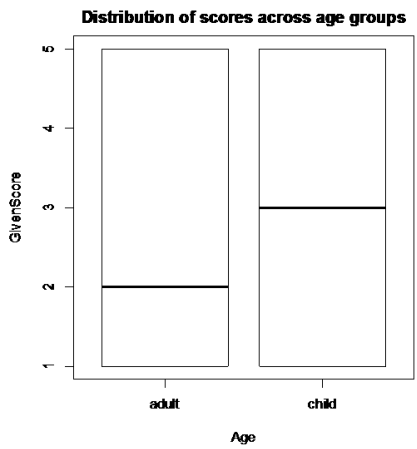


Figure 2: Impact of Age

Figure 3: Impact of Gender

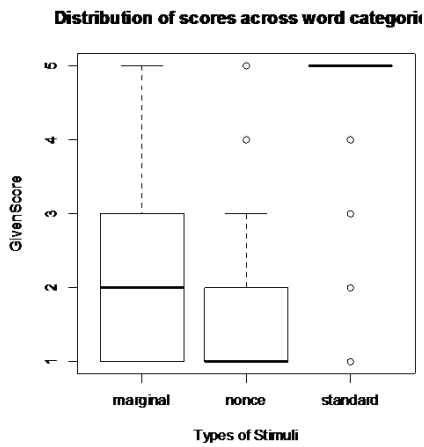


Figure 4: Impact of Word type

The goal of the statistical analysis is to determine and evaluate the strength of the correlation between the Acceptability Score and independent variables, or factorial predictor variables. The null hypothesis is that there are no statistically significant correlations among the variables. The alternative hypothesis is that such correlations do exist.

Now we will present five statistical models in turn, highlighting the advantages and the outcome of each model, and then we will discuss the overall results.

4.1. Model 1: Analysis of variance (ANOVA: parametric test for interval data)

The first model we applied to our data is ANOVA, which stands for 'analysis of variance' and is a parametric test suitable for interval data. ANOVA separates the total variation among scores into two groups: the within-groups variation, where the variance is due to chance vs. the between-groups variation, where the variance is due to both chance and the effect (if there is any). The *F* score of the ANOVA test is a ratio of the between-groups variation divided by the within-groups variation. Only when F is greater than 1, meaning that the between-groups variation is dominant, do we register an effect (King and Minium, 2008: 342–343). According to this model, the only factor that has significant impact on the distribution of scores is WordType. In our study, the difference between the distributions of acceptability scores across the three classes is found to be significantly different: F= 546, df = 2, p-value < 2.2e-16. This outcome supports the idea that the three categories of words are perceived differently by speakers. Table 6 aggregates the key parameters that characterize each type of stimuli in terms of acceptability ratings (averaged across participants).

Table 6: Distribution of average scores across Standard vs. Marginal vs. Nonce stimuli

Standard Verbs	Marginal Verbs	Nonce Verbs
MAX = 5	MAX = 3.9	MAX = 1.8
MEAN = 4.9	**MEAN = 2.4**	**MEAN = 1.5**
MIN = 4.5	MIN = 1.4	MIN = 1.2
stand dev = 0.127	stand dev = 0.551	stand dev = 0.157
variance = 0.016	variance = 0.304	variance = 0.025

We visualize these parameters in the boxplot in Figure 5. The three types of stimuli (Standard, Marginal, and Nonce) are located along the horizontal axis. The vertical axis reflects the distribution of scores.

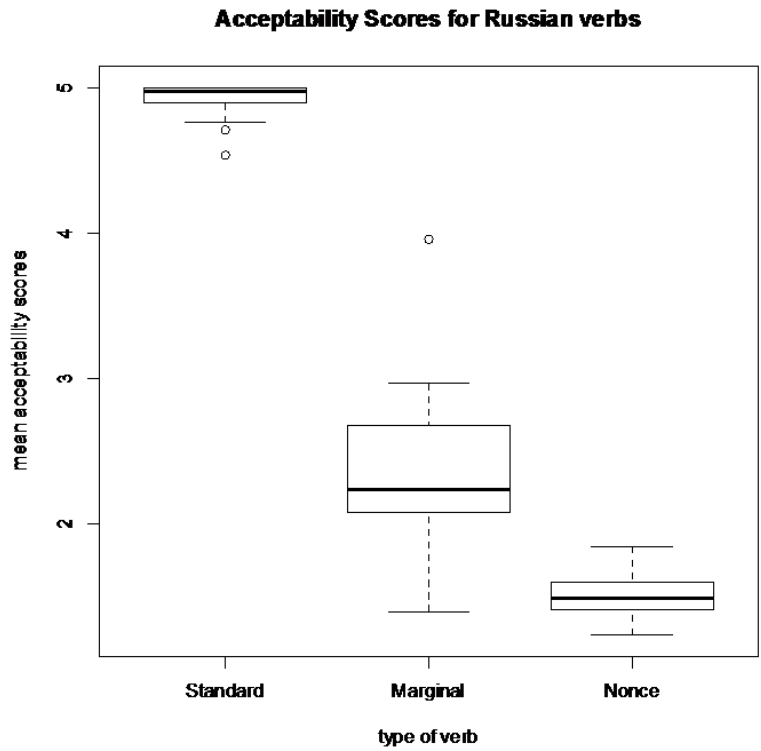


Figure 5: Three types of stimuli and distribution of acceptability ratings

Comparison of the MEAN values boldfaced in Table 6 and median values represented as thick horizontal lines within the rectangles in the boxplot suggests another crucial result. In terms of acceptability ratings, marginal words are evaluated by speakers more like nonce words rather than like standard 'normal' words.

Third, the values of variance in Table 6 show that marginal stimuli demonstrate a much larger extent of variation in terms of received scores (variance = 0.304) as opposed to standard (variance = 0.016) and nonce (variance = 0.025) stimuli. This means that marginal stimuli trigger more diversified attitudes and in this regard constitute a category on their own.

4.2. Model 2: Ordinal Logistic Regression (parametric test for ordinal data)

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). However, as Baayen (2008: 208) points out, 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, where the score '5' is higher than the score '4', the score '4' is higher than '3', and so on. For ordered values of a dependent variable, it is appropriate to use a kind of logistic regression which is **specifically designed for ordinal data** analysis – **an Ordinal Logistic Regression** (Baayen, 2008: 208–214).[5] In this analysis we used the packages languageR, rms, and MASS and the function lrm().[6] The analysis was conducted using R version 2.15.0.

The Ordinal Logistic Regression model treats the dependent variable Score as ordinal data. We explored the impact of four predicting factors – AgeGroup, Prefix, WordType, and Gender. 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 as statistically significant predictors of acceptability scores[7] – WordType and AgeGroup (with *p*-values <0.0001, or ***), and Prefix (with *p*-value = 0.0195, or *).[8] Note that here we account for main effects only. The ANOVA table (7) details the following characteristics of the significant predictors:

Table 7: Outcome of the Ordinal Logistic Regression: Wald Statistics

Factor	Chi-Square	Degrees of freedom	*p*-value
AgeGroup	59.28	1	< 0.0001
Prefix	5.45	1	0.0195
WordType	3415.95	2	< 0.0001
TOTAL	3425.06	4	< 0.0001

The *p*-value for the factor Prefix is 0.02, which is less than 0.05.[9] This means that the impact of Prefix should be considered significant, even though its significance is less than that of the factors WordType and AgeGroup which have *p*-value < 0.0001.

Comparison of the chi-square value of WordType (3415.95) with chi-square values of AgeGroup (59.28) and Prefix (5.45) in Table 7 indicates that WordType accounts for most of data, while the other two factors are very minor.

The summary of the Logistic Regression Analysis in Table 8 provides measures of predictive strength of the model. All three important measures – C,[10] Somer's Dxy,[11] and the R2 index (Harrel, 2001: 248; Baayen, 2008: 204) – are high and indicate the high predictivity of the model.

The first four lines of Table 9 represent four intercepts specific to the Ordinal Logistic Regression model. The first intercept (y>=two) contrasts datapoints with a score of 'one' to all other datapoints on the ordered scale. The second intercept (y>=three) contrasts the datapoints with the score of 'one' or

'two' on the one hand with all the datapoints with other scores, etc. We see that the four intercepts steadily decrease, reflecting the ordered scale of scores.[12]

Table 8: Outcome of the Ordinal Logistic Regression

		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 9: Coefficients

Factor	Coef	S.E.	Wald Z	Pr(>|Z|)
y>=two	0.4870	0.0586	8.31	<0.0001
y>=three	-0.4893	0.0585	-8.37	<0.0001
y>=four	-1.7137	0.0651	-26.31	<0.0001
y>=five	-2.8866	0.0830	-34.80	<0.0001
AgeGroup=child	0.4177	0.0543	7.70	<0.0001
Prefix=u	-0.1236	0.0529	-2.34	0.0195
WordType=nonce	-1.3533	0.0571	-23.70	<0.0001
WordType=standard	5.5222	0.1124	49.15	<0.0001

The four bottom lines of Table 9 indicate coefficients for factorial predictor variables with respect to a zero point, namely how O-prefixed marginal change-of-state verbs are rated by adults. The biggest difference from this zero value is found for standard stimuli (coefficient 5.5222 at the bottom line of Table 9), followed by nonce stimuli (coefficient −1.3533). A smaller difference is indicated by the coefficient 0.4177 for AgeGroup = child, and the smallest difference among the three predictors is for Prefix (coefficient −0.1236 is the least different from 0). Overall, this is in accordance with the Chi-Square values in Table 7. Note that positive values of predictor variables in Table 9 (WordType = standard and AgeGroup = child) indicate the likelihood of higher score ratings, and negative values of predictor variables (Prefix = u and WordType = nonce) correspond to likelihood of lower score ratings.

Summing up, in the Ordinal Logistic Regression analysis we approached 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 *).

4.3. Model 3: Mixed-Effects Regression Model for Ordinal Data (Parametric test for ordinal data)

The Ordinal Logistic Regression model presented in the previous section accounts for the fixed effect factors, namely WordType, AgeGroup, Prefix, and Gender. However, apart from these factors, the experimental data can also be affected by random effects factors, such as the bias of individual subjects and individual stimuli.

Figures 6–8 visualize variation across individual stimuli (standard, marginal, and nonce verbs) in terms of acceptability scores assigned to them by children and adults. In each figure, the vertical axis represents the percentage of the total possible score received by each stimulus in the experiment. As opposed to Figures 6 and 8, Figure 7 shows a greater variation in terms of ratings, and this is characteristic of marginal stimuli.

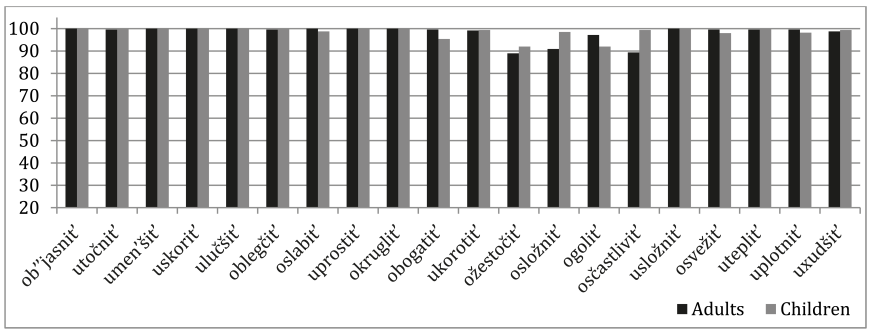


Figure 6: Variation across individual stimuli: standard verbs

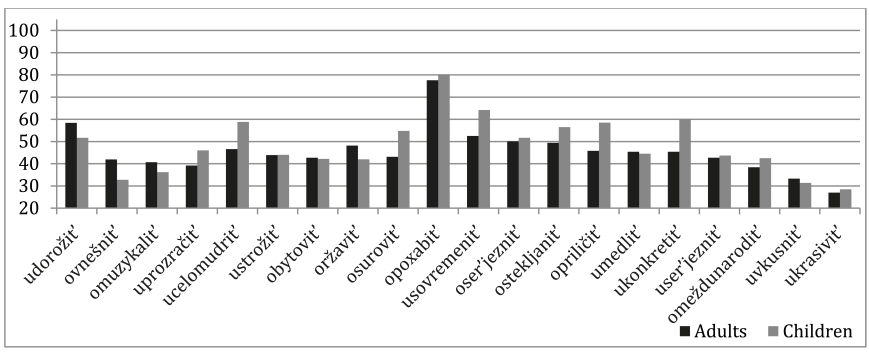


Figure 7: Variation across individual stimuli: marginal verbs

We observe high variation not only across individual marginal verbs but also across subjects: different subjects provide very different, sometimes

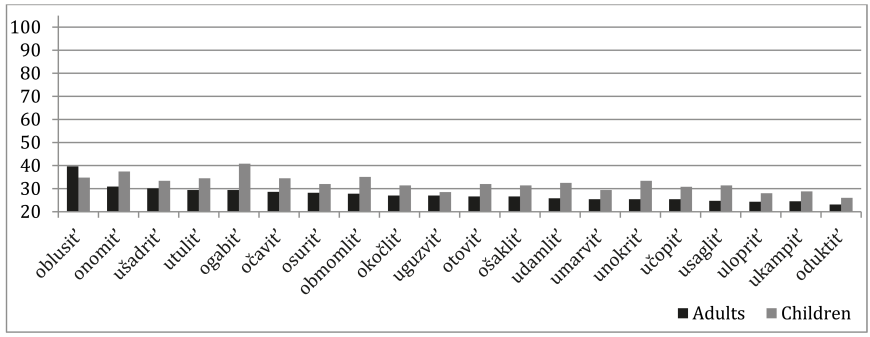


Figure 8: Variation across individual stimuli: nonce verbs

contradictory judgments for the same marginal words. Table 10 demonstrates that the verb *usovremenit'* 'modernize' was rated as a normal Russian word by 22 participants and as a non-existing word by 28. Many participants rated this word between these two extremes: 26 participants decided that it is 'a normal but rarely used word', 27 subjects suggested that 'this word sounds strange, but someone might use it', and 18 subjects evaluated it as 'a strange word unlikely to be used'. Similarly, the marginal verb *opriličit'* 'make decent' also received diverse and conflicting acceptability judgments.

Table 10: Variation across subjects regarding the same marginal stimuli

Marginal change-of-state verb	English 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

The two examples in Table 10 are representative of the distribution of scores for marginal stimuli: marginal verbs tend to trigger non-homogeneous acceptability judgments, and speakers' attitudes to such words vary (recall the high variance for marginal verbs reported in Table 6).

Both subjects ($N = 121$) and stimuli ($N = 60$) were sampled from the overall population of speakers and words, but we want to obtain a generalization about the data that would go beyond these specific subjects and specific stimuli. In other words, we need a model that can generalize over the bias of individual subjects and stimuli and determine a tendency not accounted for by random effects.

Fixed effects factors:	*Random effects factors:*
WordType: standard, marginal, nonce	Subject: 121 persons
AgeGroup: child, adult	Stimulus: 60 verbs
Prefix: O-, U-	
Gender: male, female	

A common model for experimental data where multiple subjects respond to multiple items is a Mixed-Effects Model (Baayen, 2008: 242–302). However, mixed-effects models are primarily used to explore data with nominal binomial dependent variables (0/1, A/B) (e.g. Tagliamonte and 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 (Christensen, 2015) in its version 2013.9–13 available in R version 3.0.2. We used the function clmm() which can handle the crossed random-effects structure of two factors – Subject and Stimulus.[13] It is worth mentioning that technically the Regression Mixed-Effects Model is a parametric model, but it does not assume a normal distribution for the response variable (cf. Christensen and Brockhoff, 2013; Grilli and Rampichini, 2012 for details). In this sense, it does not make parametric assumptions about the data. The Mixed-Effects Regression Model for ordinal data that we employ belongs to a family of cumulative link mixed models (CLMMs). These models use 'regression methods similar to linear models while respecting the categorical nature of the observations' (Christensen and Brockhoff, 2013: 58). As Christensen and Brockhoff (2013: 59) state, 'conceptually this is an extension of linear mixed model to ordinal observations, but computationally this model class turns out to be much more complicated. Model specification and interpretation also turn out to be more complex partly due to the discrete nature of the observations and partly due to the fact that the model is non-linear in its parameters.' Nonlinearity of the link function (Christensen and Brockhoff, 2013: 59) distinguishes this model from linear models like regression and ANOVA.

In this analysis, the factors Gender and Prefix were found insignificant in terms of predicting the dependent variable Score. After elimination of these factors, the most optimal fitted model[14] indicated significant effects for two factors: WordType and AgeGroup. Again, we take into account only main effects. Tables 11 and 12 report on the model's output regarding the random and fixed effects factors.

Summing up the outcome of the Mixed Effects Ordinal Regression Model, the impact of only two fixed effects factors was found statistically significant:

Table 11: Random-effects factors

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

Table 12: Fixed-effects factors: Coefficients

	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	***

WordType and AgeGroup. The effect of WordType is more significant than that of AgeGroup. Note that the impact of Prefix, which was less significant (*) in the Ordinal Logistic Regression, is found insignificant in the Mixed Effects model.

4.4. Models 4 and 5: Classification and Regression Trees (CART) and Random Forests (Non-parametric test for ordinal and categorical data)

Classification and Regression Trees (also abbreviated as CART) is a new method that is quickly gaining popularity in genetics, medicine, social sciences, and linguistics (Strobl *et al.*, 2009: 324; cf. recent applications in Tagliamonte and Baayen, 2012 and Baayen *et al.*, 2013).

Classification and Regression Trees is a non-parametric statistical technique that is appropriate for non-interval data. In particular, 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 which are not identified by parametric techniques (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 assume linearity and are often more flexible in modeling combinations of predictors (Faraway, 2006: v).

Because the CART model does not assume a normal distribution for the response variable (as opposed to the logistic regression model), CART can cope with a variety of data structures and types and is recommended for

unbalanced datasets. Robust results of this method are achieved by the use of recursive partitioning, bootstrapping, bagging, and cross-validation (cf. Strobl *et al.*, 2009 for details).

Apart from the high processing capacity to handle a large number of predictors non-linearly, CART analysis also offers measures of variable importance, or predictive strength of tested variables. Variable importance ranking is available via the extension of the CART method to the so-called Random Forest approach. A Random Forest is an ensemble of Classification or Regression Trees, which produces a scale of variable importance. The scale makes it possible to compare all tested predictors with each other in terms of their strength.

CART is an algorithm-based method (Faraway, 2006: 253). The outcome of a CART analysis is a graphically plotted 'tree' created via a recursive partitioning of data. The Tree represents an algorithm of data partitioning which consists of recursive binary splits, each based on one variable. The Tree outlines a decision procedure for predicting the values of the dependent variable. As a result, recursive splits subdivide the entire data set into several non-overlapping subsets of data. Each split reduces the error and increases the 'purity' of a subset of data points (the 'principle of impurity reduction', cf. Strobl *et al.*, 2009: 326). The Tree is optimal at each split. However, each local split is not necessarily globally optimal, meaning that a factor that might have a significant effect locally in the Tree, might be insignificant with regard to the entire data set.

Both Classification Tree (henceforth Ctree) and Regression Tree (henceforth Rtree) employ recursive partitioning but differ in terms of the types of response data the Tree is used for. The difference lies in the nature of the response variable: a Ctree applies to factorial dependent variables and treats the values of a dependent variable as a categorical scale, while an Rtree applies to numerical (ordinal) dependent variables (Baayen, 2008: 148). Furthermore, because Ctrees and Rtrees handle different kinds of data, they differ in mechanisms for partitioning data. A Ctree makes splits according to the principle of increasing purity of a node: after each split the subgroups of data observations should become purer, each consisting of 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, an Rtree also computes the mean within each partition.

We used both Ctree and Rtree to model the experimental data. This was both useful from the methodological perspective and reasonable in the light of uncertainty about the status of acceptability scores in terms of the type of scale that they represent (cf. Sections 1–2). In what follows we discuss the two analyses in parallel starting from Rtree (Model 4) and comparing it to Ctree (Model 5).[15]

The two resulting Trees are very similar though not entirely identical. The Rtree of acceptability ratings presented in Figure 9 approaches scores as numerical ordinally scaled data: from 5 points to 1 point. By contrast, the Ctree of acceptability ratings in Figure 10 treats the values of the dependent variable Score as categorical data: A = score '5'; B = score '4'; C = score '3', D = score '2', and E = score '1'.

CART diagrams can be understood as a set of paths defining schemas that describe the distribution of the data (Kapatsinski, 2013: 127–129). Each path begins at the top node of the diagram and proceeds down to a terminal node, or leaf, representing a distinct combination of factor values and the distribution of outcomes associated with those values. For example, in the Regression Tree in Figure 9, one path leads from node 1, which splits the data according to WordType, through node 2, which splits the data according to Prefix, down to node 3. This path shows the outcome of scores for standard verbs with the prefix U-, where nearly all the scores are 5, with outliers at the scores of 4 and 3. The same combination is shown in the same path in the Classification Tree in Figure 10, going from node 1 through node 11, to node 12, where we get a different graphic representation of the same scores.

Although the Ctree expands to the left, while Rtree stretches to the right, they make almost identical splits, just in different order. Crucially, both Trees demonstrate that WordType determines the major split of data at the root node (node 1), followed by Prefix at the second level, and AgeGroup at the third level.

The root node is the same in both Trees – WordType. Note that the decision rule of the root node partitions data into two large subsets, grouping together marginal and nonce verbs and setting them apart from standard verbs. We can interpret this partitioning as indicating a close connection between marginal and nonce verbs in terms of their similar acceptability ratings and a larger distance between marginal verbs and standard verbs. Recall that this generalization is also supported by the ANOVA analysis: marginal verbs as a group pattern more similarly to nonce verbs than to standard verbs.

In both trees standard verbs are further split according to Prefix. Terminal (leaf) nodes 12 and 13 of the Ctree (Figure 10) and nodes 3 and 4 of the Rtree (Figure 9) demonstrate that standard verbs prefixed in U- as a group receive slightly higher acceptability ratings (i.e. are better accepted) than standard verbs prefixed in O-. In particular, the plots of terminal nodes show that among verbs prefixed in O- there are more outliers that receive scores lower than '5' than in the group of U-verbs. This is supported by a total of 2,420 data points (see the numbers that appear on the terminal nodes).

In both trees, in the branch opposite standard verbs, WordType further determines the split into marginal and nonce stimuli. In the Ctree, marginal

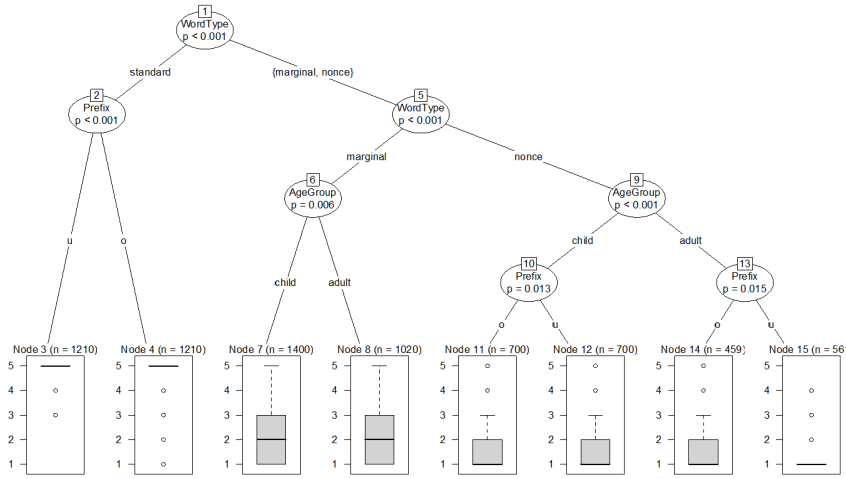


Figure 9: Regression tree of acceptability ratings: scores are treated as numerical ordinal data – from 5 points to 1 point

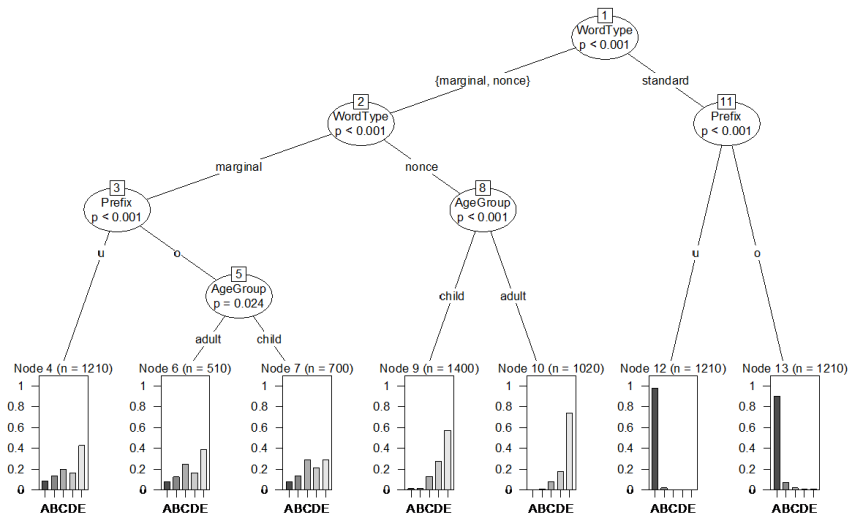


Figure 10: Classification tree of acceptability ratings: scores are treated as categorical data: A-score '5'; B-score '4'; C-score '3', D-score '2', E-score '1'

verbs are further subdivided according to Prefix and AgeGroup. These subsequent splits suggest that marginal verbs prefixed in U- (node 4 in Figure 10) receive slightly more rejections (score '1') than O-verbs (nodes 6 and 7). Meanwhile, for O-verbs we can observe an interaction effect of Prefix and AgeGroup:

adult speakers tend to reject marginal verbs prefixed in O- more often than children (compare the bars representing 'E' score in nodes 6 and 7). The Ctree suggests an AgeGroup effect for nonce verbs as well (node 8 in Figure 10). Again, adults tend to completely reject nonce verbs regardless of their prefix more often than children do (compare 'E' bars in terminal nodes 9 and 10).

The Rtree has the same predictors in the marginal and nonce branch, slightly rearranging their order of application. Crucially, marginal verbs are partitioned exclusively according to the factor of AgeGroup, but the difference between adults and children in this domain must be very small because nodes 7 and 8 (Figure 9) look identical. The group of nonce verbs, by contrast, is affected by the interaction of AgeGroup and Prefix: for children both O- and U-verbs pattern pretty much the same (compare the nodes 11 and 12), while for adults nonce verbs prefixed in O- (as opposed to U-) tend to be more acceptable and more diverse in terms of their ratings and include more outliers with scores higher than '1'.

Summing up, both trees show high-level interactions of WordType, AgeGroup, and Prefix. Both Ctree and Rtree visualize data distribution with respect to three factors. The structure of both trees is largely similar: 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. The role of these factors in the overall data distribution is different (much smaller), as clearly shown in the Random Forest analysis.

In order to compare the two Random Forest analyses consider Figures 11 and 12.

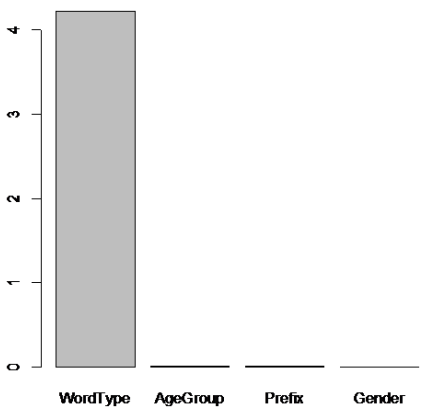


Figure 11: Variable importance scale for ordinal data (5>4>3>2>1) modelled in Rtree

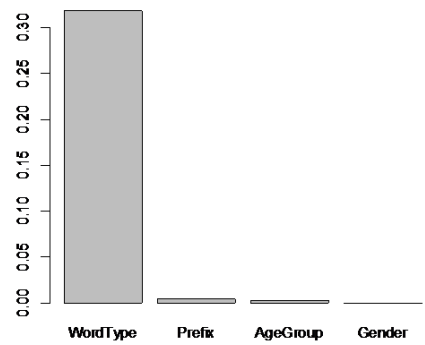


Figure 12: Variable importance scale for categorical data (A, B, C, D, E) modelled in Ctree

Both figures present barplots of variable importance scores for factorial predictors of acceptability ratings. Figure 11 presents the outcome of the Random Forest analysis of acceptability scores taken as ordinal data (5>4>3>2>1), while Figure 12 is the result of the Random Forest analysis of acceptability scores taken as categorical data (A, B, C, D, E).

Both barplots visualize a scale of relative importance, where the predictors of the dependent variable Score are ranked according to their relative strength. Each bar represents one predictor. Both plots depict the same four factors and arrange them almost identically. First of all, both plots show that WordType is by far the strongest predictor, while the impact of other factors is close to zero. Both plots show that Gender is the weakest predictor of all (recall that it appeared in neither of the Trees). Prefix and AgeGroup are ranked differently: Forest analysis of categorical data (Figure 12) suggests that Prefix is slightly stronger than AgeGroup, while Forest analysis of ordinal data (Figure 11) 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.

5. Discussion

The goal of this section is to summarize the outcomes of various statistical techniques and highlight what was consistent throughout all analyses.

We find that the outcomes of different models are similar. In the Ordinal Logistic Regression Model we found a significant effect for all three factors. In the Mixed-Effects Regression Model for Ordinal data only WordType and AgeGroup showed a significant effect. In the Trees and Forests analysis only WordType was the major predictor while AgeGroup and Prefix gain significance within local subsets of data.

We suggest that the latter model is the most insightful and fruitful regarding this data. As a non-parametric test, CART 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 predictive power from the perspective of 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 minor, as shown on the variable importance scale. Indeed, this effect is revealed in subtle interactions of the factors depicted in the Classification and Regression trees.

Our research questions asked: (1) whether the prefix O- (which is most productive in Russian) is preferred over the prefix U-; (2) whether adults have

more conservative judgments than children; and (3) whether marginal verbs are perceived to be more like standard or more like nonce verbs.

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 Mixed-Effects Regression Model for ordinal data and ANOVA, where Prefix is not found to be significant at all. In both Random Forest analysis and Ordinal Logistic Regression, our hypothesis was supported: the prefix O- did garner higher scores than the prefix U-, though the effect was not strong.

The low predictive strength of AgeGroup revealed by Random Forest corresponds to what was found by ANOVA. At the same time, this contradicts the result of the Ordinal Logistic Regression and the Mixed-Effects Regression analyses, where the effect of AgeGroup was found to be statistically significant, but less so than both Prefix and WordType. Both of the regression analyses showed the effect predicted by our hypothesis, namely, that adults were less likely to give high scores to marginal words than children, though this effect was even more marginal than that of Prefix.

The major role of WordType is supported by all models that we applied and showed that marginal words were perceived to be much more like nonce words than like standard words.

The effect of Gender is insignificant according to all models where it was tested.

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 to Russia, which implies strong linguistic norms and in particular strong concern for the 'purity' of the literary language. Note that marginal words are semantically transparent, while nonce words are not. Thus, our finding that marginal words are rated more like nonce words than like standard words indicates that speakers are more sensitive to frequency than to semantic transparency. This suggests that frequency, which is related to performance, is a stronger factor than competence (ability to unpack morphological patterns). Therefore, memory may be a stronger factor than the use of productive rules. On the other hand, marginal words exist on their own terms and differ from both standard and nonce words in terms of much higher variation across stimuli.

6. Conclusions

In this article we report on an experimental study that targeted marginal change-of-state verbs in Modern Russian. We tested whether the prefix (O- vs. U-), gender and age of speakers, and word type correlate with higher or lower acceptability of words in the perception of native speakers. Both prefix and age factors behaved according to our predictions, and we discovered that marginal

words are perceived to be more like nonce words than like standard words. We approached the data from different perspectives, applying both parametric and non-parametric statistical tests, including models specifically designed for handling ordinal and categorical data. We arrive at conclusions that can be summarized in four key points.

First, we argue that all five models that we applied are appropriate for this data set to a greater or lesser degree. Nevertheless, the five models treat this data set differently. ANOVA is a parametric model that assumes interval level of data measurement, the interpretation supported in our case by the numeric scores of 1 to 5 points. Ordinal Logistic Regression and Mixed-Effects Regression Models are two parametric models specifically designed for handling ordinal data, thus avoiding any assumption of intervalness. Indeed, ordinal scale is a possible interpretation for our set of scores, since our scores are values that are internally ordered but do not necessarily have intervals of equal magnitude. A Regression tree is another model suitable for a numerical ordinal variable and might be even more trustworthy due to its non-parametric nature (e.g. no assumption of a normal distribution). Finally, a Classification tree interprets the scores that participants assigned as a set of categorical values, which corresponds to our set of descriptive evaluative statements that accompanies acceptability scores.

Second, we observe that parametric tests provide outcomes comparable with non-parametric models. The five models focus on different aspects of data, but all models identify WordType as the major predictor. The differences concern the factors AgeGroup and Prefix that are of marginal importance.

Third, we advocate Classification Tree combined with Random Forests as the most conservative model that is appropriate for this data set. This model is non-parametric and is designed for categorical dependent variables. Moreover, we suggest that this model is most informative regarding the marginal verbs that are the focus of this study. In particular, the Ctree 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 minor overall predictive power from the perspective of 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 Ctree and the highest bar in the Random Forest plot. To be precise, the outcome of Random Forest analysis indicates that AgeGroup and Prefix do have some importance but their effect is very minor. This effect is revealed in the interactions of the factors. Another advantage of CART is that this technique is user-friendly and produces visualizations that are relatively easy to read and interpret.

Fourth, in this study we propose that the use of a culturally entrenched grading scale is an advantage in an experimental design. In our experiment we

used evaluative judgments aligned with the numeric scale of five points commonly used in Russian school and university grades. The scale of five points is highly culturally embedded, familiar, and all subjects can rely on it.

We find it not entirely wrong to: (1) translate the acceptability scores into numeric values; and (2) apply parametric statistics like analysis of variance (ANOVA) to this data. Because the data collected via this type of scale can be interpreted in terms of different levels of data measurement, comparison of the outcomes of parametric and non-parametric statistical models designed for handling different types of data is of key importance and bears implications for similar studies.

Linguistics has undergone a quantitative turn, establishing new standards for data analysis (Janda, 2013). The major contribution of this article consists of detailed applications of several statistical models documented in R scripts that can be used by linguists and possibly by scholars of other fields that work with Likert and Likert-type scale data. Whereas in the past, the prevailing concern in analyzing such data was the need to adhere to the assumptions of a narrow set of statistical models, the variety of statistical models available today supports the choice of an experimental design driven by specific research questions and the sociocultural background of participants.

About the authors

Anna Endresen is a cognitive linguist and a member of the CLEAR research group (Cognitive Linguistics: Empirical Approaches to Russian) at UiT The Arctic University of Norway, where she completed PhD studies and later worked as a university lecturer/associate professor.

Laura A. Janda is also a member of the CLEAR research group and Professor of Russian Linguistics at UiT The Arctic University of Norway.

Notes

1. We use symbols >>> and > to indicate greater and minor differences in predictive power of factors. For example, according to the Ordinal Logistic Regression Model, the importance of WordType factor is much larger than that of AgeGroup; and the effect of AgeGroup is somewhat larger than that of Prefix.

2. For the history of conflicting views see Gardner (1975); a brief summary is given in Knapp (1990).

3. In using the term *acceptability judgments* instead of *grammaticality judgments* we follow Bermel and Knittl (2012).

4. Note that in *ob"jasnit'* 'clarify' as well as in *oblegčit'* 'simplify, lighten' we deal with OB-, a phonologically conditioned allomorph of the prefix O-.

5. 'When a data frame is read into R, the levels of any factor are assumed to be unordered by default' (Baayen, 2008: 209). Therefore, 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")).

6. What is crucial for the function lrm() of the Ordinal Logistic Regression model is that it 'assumes that the effects of our predictors <…> are the same <…> across all levels of our ordered factor' (Baayen, 2008: 212). Although this might somewhat simplify the outcome, we nevertheless obtain an important generalization about the statistically significant predictors of data distribution.

7. The formula that we used: dat.lrm2 = lrm(Score ~ AgeGroup + Prefix + WordType, data = dat, x = T, y = T) and the command used was anova(dat.lrm2).

8. According to the common set of codes that indicate significance, the number of stars corresponds to the level of significance: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

9. 'In language and linguistic research it is customary to take an *alpha* decision level of 5 percent ($p < 0.05$). This means that there is less than 5 percent probability that rejecting the null hypothesis will be an error' (Cantos Gómez, 2013: 49; cf. also Baayen, 2008: 188).

10. 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.

11. 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)'.

12. Although the coefficients of the intercepts are not designed to be a measure of intervalness, it is interesting to note that the intervals between them (0.96, 1.22, and 1.18) are indeed roughly equal.

13. We are indebted to Rune Haubo Bojesen Christensen for pointing out this possibility.

14. The formula used is: fm2 <- clmm(Score ~ AgeGroup + WordType + (1|Stimulus) + (1|SubjectCode), data=dat, Hess=TRUE).

15. Both analyses were carried out in R version 2.15.0.

References

Baayen, R. H. (2008). *Analysing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., Janda, L. A., Nesset, T., Endresen, A., and Makarova, A. (2013). Making choices in Slavic: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37: 253–291. https://doi.org/10.1007/s11185-013-9118-6

Bermel, N. and Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8 (2): 241–275. https://doi.org/10.1515/cllt-2012-0010

Blaikie, N. (2003). *Analyzing Qualitative Data*. London: SAGE Publications Ltd. https://doi.org/10.4135/9781849208604

Cantos Gómez, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield: Equinox Publishing.

Christensen, R. H. B. (2015). *Ordinal – Regression Models for Ordinal Data*. R package version 2015.6-28. Software and manual retrieved on 19 May 2017 from https://cran.r-project.org/web/packages/ordinal/index.html

Christensen, R. H. B. and Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de la Société Française de Statistique* 154 (3): 58–79.

Cohen, L., Manion, L., and Morrison, K. (2000). *Research Methods in Education*, 5th ed. London: Routledge Falmer. https://doi.org/10.4324/9780203224342

Collins, C., Guitard, S. N., and Wood, J. (2009). Imposters: An online survey of grammaticality judgments. *NYU Working Papers in Linguistics 2: Papers in Syntax*. Retrieved on 19 May 2017 from http://linguistics.as.nyu.edu/docs/CP/2345/collins_guitard_wood_imposters_online_09_nyuwpl2.pdf

Dąbrowska, E. (2010). Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27: 1–23. https://doi.org/10.1515/tlir.2010.001

Dubois, D. (2013) Statistical reasoning with set-valued information: Ontic vs. epistemic views. In C. Borgelt, Gil, M. A., Sousa, J. M. C., and Verleysen, M. (Eds) *Towards Advanced Data Analysis by Combining Soft Computing and Statistics. Studies in Fuzziness and Soft Computing* 285: 119–137. Berlin/Heidelberg: Springer-Verlag.

Endresen, A. (2013). Samostojatel'nye morfemy ili pozicionnye varianty? Morfologičeskij status russkix pristavok *o-* i *ob-* v svete novyx dannyx: korpus i èksperiment [Distinct morphemes or positional variants? Morphological status of the Russian prefixes *o-* and *ob-* in the light of new evidence: corpus and experiment]. *Voprosy jazykoznanija* 6: 33–69.

Endresen, A. (2014). *Non-Standard Allomorphy in Russian Prefixes: Corpus, Experimental, and Statistical Exploration*. Doctoral dissertation. University of Tromsø: The Arctic University of Norway. Retrieved on 19 May 2017 from http://hdl.handle.net/10037/7098

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton, FL: Chapman and Hall/CRC.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research* 45: 43–57. https://doi.org/10.3102/00346543045001043

Grilli, L. and Rampichini, C. (2012). Multilevel models for ordinal data. In R. S. Kenett and S. Salini (Eds) *Modern Analysis of Customer Surveys: with Applications using R*, 391–408. Chichester: John Wiley and Sons.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer series in Statistics. https://doi.org/10.1007/978-1-4757-3462-1

Haspelmath, M. (2002). *Understanding Morphology*. London: Oxford University Press.

Jaccard, J. and and Wan, C. K. (1996). *LISREL Approaches to Interaction Effects in Multiple Regression.* Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781412984782

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education* 38: 1212–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Janda, L. A. (Ed.) (2013). *Cognitive Linguistics: The Quantitative Turn. The Essential Reader*. Berlin and Boston, MA: De Gruyter Mouton.

Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* 89: 110–148. https://doi.org/10.1353/lan.2013.0003

Keller, F. and Asudeh, A. (2001) Constraints on linguistic coreference: Structural vs. pragmatic factors. In J. D. Moore and Stenning, K. (Eds) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* 483–488. Mahawah, NJ: Lawrence Erlbaum Associates.

Kim, J.-O. (1975). Multivariate analysis of ordinal variables. *American Journal of Sociology* 81: 261–298. https://doi.org/10.1086/226074

King, B. M. and Minium, E. W. (2008). *Statistical Reasoning in the Behavioral Sciences.* Hoboken, NJ: Wiley.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research* 39 (2): 121–122. https://doi.org/10.1097/00006199-199003000-00019

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces* 46: 151–160. https://doi.org/10.2307/2574595

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review* 35: 515–524. https://doi.org/10.2307/2092993

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781412963947

Likert, R. (1932). *A Technique for the Measurement of Attitude*s. Doctoral dissertation. Columbia University. Series *Archives of Psychology* 22: 5–55. NY: The Science Press. Retrieved on 19 May 2017 from http://www.voteview.com/pdf/Likert_1932.pdf

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education* 39 (9): 970. https://doi.org/10.1111/j.1365-2929.2005.02237.x

R Development Core Team. (2010). *R: A Language and Environment for Statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Rietveld, T. and van Hout, R. (2005). *Statistics in Language Research: Analysis of Variance.* Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110877809

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago, IL and London: The University of Chicago Press.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14 (4): 323–348. https://doi.org/10.1037/a0016973

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24 (2): 135–178. https://doi.org/10.1017/S0954394512000129

Townsend, Ch. E. (1968). *Russian Word-Formation.* Bloomington, IN: Slavica Publishers. Reprint edition from 2008.