Ordinal GAMMs: a new window on human ratings

Divjak, Dagmar; Baayen, Harald

Ordinal GAMMs: a new window on human ratings

R. Harald Baayen and Dagmar Divjak

abstract

The vast majority of linguistic theories are built on a peculiar type of data: acceptability or grammaticality ratings. Traditionally these ratings were obtained through introspection by the analyst, an approach that is problematic in many (if not most) respects. Linguists addressed (part of) the issue by starting to elicit ratings from largish numbers of native speakers. Yet, this caused a new problem: due to the unpopularity of ordinal data in disciplines that drive the development of statistical analysis, few techniques are available that handle this type of date with grace. In our contribution, we explain how Generalized Additive Mixed Models can be used to explore ordinal data in all its complexity using the `mgcv` package in `R`.

1 Introduction

Frequency is among the most robust predictors of human performance (Hasher and Zacks, 1984). A large number of studies have investigated the extent to which different forms of linguistic behaviour would be frequency-driven, and evidence has been found for a range of phenomena, from processing single words to acquiring knowledge of the sets of verbs that are used in complex argument structure constructions (for book-length overviews see Sedlmeier and Betsch 2002; Ellis 2002; Diessel 2007; Divjak and Gries 2012; Gries and Divjak 2012). These findings have spurred on the development of usage-based linguistics, which assume that frequency, as proxy of experience, plays a central role in the emergence and entrenchment of linguistic units: surface distributions contain the necessary information to build up adequate mental linguistic representations. Surface token frequency motivates learning through repetition: the token is the instance that is repeated and subsequently learned. The more often a pattern is experienced, the easier it becomes to access and use (see articles in Bybee and Hopper 2001; for recent studies on the way discrimination learning may explain frequency effects, see Baayen 2011b,a).

Yet, frequency of occurrence does not fit seamlessly into standard linguistic practice, and this for two reasons. First, the one area where frequency seems to have run into problems is that of acceptability or grammaticality judgments, a form of data on which much linguistic theorizing relies, irrespective of tradition. In accord with usage-based theory one would expect that "grammaticality or acceptability judgments are heavily based on familiarity, that is, the speaker's experience with language in use. Sequences of linguistic units that are of high frequency or resemble sequences of high frequency will be judged more acceptable than those that are of low frequency or do not resemble frequently used structures" (Bybee and Eddington (2006, p. 349), see also Bannard and Matthews 2008; Shaoul et al. 2013; Arnon and Snider 2010 for frequency effects for word sequences across a variety of tasks). A number of studies in both the generative and usage-based traditions have, however, confirmed the existence of a grammaticality-frequency discrepancy, if not a "gap" (Kempen and Harbusch, 2005, 2008), for acceptability ratings: corpus frequencies are poor predictors for off-line acceptability ratings, in particular at the lower end of the frequency spectrum, in both morphology and syntax (Keller 2003; Kempen and Harbusch 2005, 2008; Arppe and Järvikivi 2007; Divjak 2008; Bader and Häussler 2010; Bermel and Knittl 2012b,a; but see the opposite tendency in the results of Lapata et al. 1999 for adjective-noun combinations). This has strengthened generativists in their belief that "simple frequency data" could and should be ignored

in theoretical linguistic analyses. Because acceptability judgments provide a substantial part of the empirical foundation of dominant linguistic traditions, it is important for linguists to understand how acceptability relates to frequency. These considerations led to the study reported in Divjak (2016).

Divjak (2016) highlighted a second problem: acceptability ratings elicited on a Likert scale yield ordinal data and the analysis of the resulting ordinal data tends to be cumbersome, especially for data with crossed random effects for subjects and items.[1] One option is to ignore that the data are ordinal, and proceed with a standard regression analysis. When the processes that give rise to the ratings are underlyingly continuous and uniform, this may not make much of a difference. Regardless, two issues arise. First, we do not know whether the ratings are discretizations of a nice and smooth underlying linear scale. If this is not the case, treating an ordinal variable as simply integer-valued may yield implausible results. Second, even if the ratings reflect an underlying linear scale, the analysis may lack precision. This problem becomes acute when we consider the question of how to turn the real-valued predictions of a linear model, fitted to ordinal data, back into the original rating categories. For instance, in a linear model, ratings on a 5-point Likert scale may be paired with predicted values ranging from 0.5 to 5.1. What criteria should be used to bin the predicted values? Where should the boundaries between the bins be posited? Here, many choices can be made, and which choice is made will influence prediction accuracy — unfortunately, in an ad hoc and unprincipled way.

Those wishing to travel the royal road will quickly discover that ordinal regression is indeed not straightforward. For instance, including more than one random-effect factor may be problematic due to limitations on available packages in freely available software such as R. Furthermore, software tends to assume covariates have a linear effect, which is a strong assumption that does not always hold. Examples of the consequences of (ignoring) non-linearities for understanding experimental data are discussed in detail in Baayen et al. (2017) in the context of the generalized additive model (Wood, 2006, 2011). Fortunately, recently, Wood and colleagues extended their `mgcv` package with further algorithms (Wood et al., 2016), one of which makes it possible to analyse ordinal categorical response variables properly. One or more random effect factors can be included where necessary, and effects of numeric predictors are no longer restricted to being strictly linear.

The goal of this study is to introduce this new method of analysis by means of a worked example, exploring a subset of the data presented in Divjak (2016). This study investigated the relation between frequency and acceptability using corpus- and behavioral data on the distribution of infinitival and finite that-complements in Polish. In what follows, we begin with introducing the data, after which we first present the GAM model, followed by its application to the Polish data set.

2 The data

Polish verbs exhibit substantial subordination variation and for the majority of verbs taking an infinitival complement, the that-complement occurs with low frequency (< 0.66 ipm, estimate taken from a 1.5 billion word corpus). These low-frequency that-clauses, in turn, exhibit large differences in how acceptable they are to native speakers.

An example illustrates that it is possible to use both infinitival and finite complements in co-referential sentences with *decide* as the main verb, such as (1), but not in sentences such as (2) with

[1] It is not necessary to make use of such ordinal scales, ratings on a Visual Analogue Scale have been found to be provide useful real-valued ratings (Hayes and Patterson, 1921; Freyd, 1923; Funke and Reips, 2012; Geeraert, 2016; Geeraert et al., 2017).

want as the matrix verb. The question is: how do speakers know when a that-alternative is and is not available?

(1)

Zdecydował	wyjechać
Decided PF.IND.PAST.MASC.3SG	leave PF.INF
He decided to leave.	
Zdecydował,	że wyjedzie.
Decide PF.IND.PAST.MASC.3SG	that leave PF.IND.NON_PAST.3SG
He decided that he would leave.	

(Grzegorczykowa 2006, 83)

(2)

Chciał	wyjechać
Want IMPF.IND.PAST.MASC.3SG	leave PF.INF
He wanted to leave.	
*Chciał,	że wyjedzie.
Want IMPF.IND. PAST.MASC.3SG	that leave PF.IND.NON_PAST.3SG
He wanted that he would leave.	

This phenomenon has received ample attention in research on Subject/Object Control and Subject Obviation within generative frameworks, and this for a range of languages including Polish (Bondaruk, 2004; Dziwirek, 1998, 2000; Przepiórkowski and Rosen, 2005). Control verbs in Polish differ with respect to whether they allow, require or resist the presence of a complementizer and the meaning of the verb does not affect this (Bondaruk, 2004, p. 208). This begs the question of how learners acquire part of a subordination system for which there is no apparent semantic or functional motivation.

In Divjak (2016), the extent to which usage, and a speaker's experience of it, contributes to the acceptability of complex lexico-syntactic structures was captured by exploring the relation between, on the one hand, off-line acceptability ratings (Sprouse, 2013) for verbs that occur with low frequency in that-constructions and, on the other hand, a range of variables capturing information relating to the (co-)occurrence, morphology, and semantics of the verbs and the that-construction. We will use part of this dataset in our contribution and recap here the design of the study.

The 285 experimental sentences were (shortened) authentic that-sentences extracted from the newspaper section within the PELCRA reference corpus of Polish (`http://korpus.ia.uni.lodz.pl/`); in case no that-sentences were attested, some were created from infinitive sentences found in the same sub-corpus using the most likely form of the that-clause, as judged by 5 native speakers of Polish. In order to neutralize lexical effects of any items other than the verb (Schütze and Sprouse, 2014, p. 39), three different lexicalizations were provided for each of the 95 verb×that-construction combinations. 25 filler and 10 benchmark sentences were adapted from authentic sentences extracted from newspapers to be comparable to the experimental sentences in plausibility, complexity and length and to instantiate grammaticality levels ranging from -2 to +2. Overall, the ratio between experimental sentences and fillers was 1:9 in the survey, and within each block of 8 sentences, only 1 was an experimental sentence. In each questionnaire, 5 of the experimental sentences (each with a different verb) and 25 fillers were randomly assigned to 5 blocks and then shuffled within blocks. The first block was preceded by a block of 5 benchmark sentences; the last block was followed by 5 benchmark sentences. Participants were asked to indicate "how Polish this sentence

sounds" on a 5-point Likert scale where the lowest score signaled "very strange, unnatural Polish", the highest score was reserved for "natural Polish and the midpoint marked "OK Polish, could be heard". Participants were ensured there were no right or wrong answers and were asked not to revisit previous answers.

Raters do not necessarily interpret a scale in exactly the same way, and do not rate items in the precisely the same fashion: some tend to assign high scores, while others err on the side of caution. To capture this subject-specific behavior, the variable "Rater Generosity" was introduced. Rater Generosity is the by-participant mean of the scores assigned to the 25 filler items, accommodated in a five-level (ordinal) category, with cut-off points that are symmetrical on either side of zero and display a normal distribution.

A wide range of variables, including measures of frequency and association, were included in Divjak (2016). Frequency data was obtained from the 1.5 billion word version of the NKJP, the Polish National Corpus. All texts, with the exception of older prose and a small number of contemporary prose texts, were created from the 1990s onwards. The corpus is fully parsed with an overall tagging accuracy of up to 98% (Adam Przepiórkowski 2007, p.c.). Data on the that-construction for the 95 verbs studied were extracted by means of regular expressions written for the stand-alone version of Poliqarp (Janus and Przepiórkowski 2005). We consider two here: Verb Frequency and the Reliance of a verb on a lexico-grammatical pattern.

The unigram frequency of the verb in the corpus, a measure of overall frequency, was included to give an idea of how likely raters would be to know the verb in question. Within the complete sample, occurrences ranged from 295 to 9330418 (in a corpus of 1.5 billion words).

Reliance (Schmid, 2000, p.56) measures the degree to which a verb depends on a lexico-grammatical pattern, relative to the occurrence of the same verb in other patterns. It is a unidirectional relative frequency measure, defined as the frequency of a verb×construction combination given the frequency of the verb. Reliance can be considered a conditional probability: Reliance is $\Pr(c|v) = \Pr(v \cap c)/\Pr(v)$. It gives an idea of how likely the construction is to follow if the verb is known. The rank list for Reliance is often topped by lexemes which are highly specialized for occurrence in the given pattern but may be fairly infrequent overall (Schmid 2010, 110).

3 The data re-analyzed

For the analyses, we need the `mgcv` (Wood, 2006) and `itsadug` (van Rij et al., 2016) packages, as well as the data itself, which is available as `polish.rda` at `https://opendata.uit.no/dataverse/trolling`, and which we have locally available in a folder named `data`.

```
library(mgcv)
library(itsadug)
load("data/polish.rda")
head(dat,3)
  AcceptabilityRating RankConstructionVerbReliance LogFrequencyVerb
1                   3                           963          10.0488
2                   3                           963          10.0488
3                   3                           963          10.0488
  RaterGenerosity Verb Subject
1            0.04    1      11
2           -0.44    1      21
3            0.40    1      41
```

The reliability measure was rank-transformed in order to avoid problems with its non-normal distribution, characterized by several outlier values. Verb frequency was log-transformed to avoid adverse

effects of outlier frequencies, after backing off from zero by adding 1 to all frequencies.

3.1 Ordinal regression with GAMMs

We are interested in the probability that the `AcceptabilityRating` (Y) takes a value from $r = 1, \ldots, R$, the r being labels for ordered categories. Wood et al. (2016), following Kneib and Fahrmeir (2006), estimate the probability $\Pr(Y = r)$ by positing $R + 1$ cutoff points α,

$$-\infty = \alpha_0 < \alpha_1, \ldots, \alpha_{R-1} < \alpha_R = \infty,$$

that partition the real axis into R intervals. A given rating that falls in category r ($y = r$) is coupled with a latent variable $u = \mu + \epsilon$ that assumes a value in the r-th rating 'bucket' $\alpha_{r-1} < u \leq \alpha_r$. Given the logistic cumulative distribution function $F(u)$ for u,

$$F(u) = \frac{e^u}{1 + e^u},$$

we obtain the following equality:

$$\begin{aligned} \Pr(Y = r) &= \Pr(\alpha_{r-1} < u \leq \alpha_r) \\ &= F(\alpha_r - \mu) - F(\alpha_{r-1} - \mu). \end{aligned}$$

In other words, the probability that $Y = r$ is assessed through the probability that u falls in the r-th rating 'bucket'.

The latent variable u is modeled as an additive function of the predictors X_i or smooth functions $s(\cdot)$ of one or more of these predictors, as in the following example, where we have a linear predictor with slope β_1 and a wiggly second predictor:

$$u_i = \beta_0 + \beta_1 x_{i1} + s(x_{i2}).$$

In this set-up, predictors can co-determine the probability in which 'bucket' u will be located, and hence what category Y is most likely to be in, given the predictors. The cut-points α are estimated alongside the model parameters of the smoothing functions $s(\cdot)$ (and for mixed models, the parameters of the random effects). It is important to note that this takes the problem of having to determine cut-off points for assigning predicted values to one of the 5 discrete ratings, mentioned above, out of the analyst's hands. For further technical details on how the model is fitted, the reader is referred to Wood et al. (2016).

To make this more concrete, consider predicting `Acceptability Rating` from `Rank Construction Verb Reliance`, `Log Frequency Verb` and `Rater Generosity`, while restricting the effects of these predictors to being strictly linear. The following model, which also contains random intercepts for `Verb`, is fitted with the `gam` function of `mgcv`, specifying through the `family` directive that the response is an ordered factor with 5 values.

```
dat.gam0 = gam(AcceptabilityRating ~
               RankConstructionVerbReliance + LogFrequencyVerb + RaterGenerosity +
               s(Verb, bs="re") ,
               data=dat, family=ocat(R=5))
```

The summary of this model is found in Table 1, which shows that the probability of a higher rating score increases for greater values of `Rank Construction Verb Reliance` and `Rater Generosity`. `Log Frequency Verb` appears to have no effect, consistent with the findings of Divjak (2016). There is no solid evidence for by-subject random intercepts, unsurprising as `Rater Generosity` represents subjects' average locations on the rating scale, based on the filler materials of the same experiment. The boundaries α that demarcate the counterparts of the acceptability rating categories on the real axis can be extracted from the model object as follows:

```
dat.gam0$family$getTheta(TRUE)
[1] -1.0000000  0.2260885  1.3579938  2.6475822
```

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-0.2323	0.5884	-0.3947	0.6931
RankConstructionVerbReliance	0.0015	0.0003	5.8312	< 0.0001
LogFrequencyVerb	-0.0366	0.0490	-0.7467	0.4553
RaterGenerosity	1.4523	0.1121	12.9550	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Verb)	71.1376	93.0000	300.5126	< 0.0001

Table 1: Model summary for the baseline GAMM fitted to the ordered categorical acceptability ratings, with linear predictors only. The term s(Verb) denotes the by-verb random intercepts.

3.2 A worked example

Having outlined the way in which the generalized additive model can be extended to handle categorical ordinal response variables, we can now harness the true power of the GAMM to delve deeper into the intricacies of acceptability ratings. Specifically, we can ask whether the effects of the three item-bound covariates (`RankConstructionVerbReliance`, `LogFrequencyVerb`, and `RaterGenerosity`) are truly strictly linear. Furthermore, might these predictors interact? Especially `RaterGenerosity`, as a measure tapping into individual differences, could well show subject-specific differentiation across the effects of `Rank Construction Verb Reliance` and `Log Frequency Verb`. The frequency and reliance measures may also show an interaction, firstly, because both tap into users' experiences with the verbs and the constructions in which they appear, and secondly, because both measures are, by their definition, mathematically related, as explained above. As a consequence, it is unlikely that their effects would be completely orthogonal.

We therefore fitted a sequence of increasingly complex models, incrementally testing for increasingly complex interactions. The first model relaxes the linearity assumption while not yet incorporating interactions:

```
dat.gam1 = gam(AcceptabilityRating ~
               s(RankConstructionVerbReliance) + s(LogFrequencyVerb) + s(RaterGenerosity) +
               s(Verb, bs="re") ,
               data=dat, family=ocat(R=5))
```

The second model introduces a nonlinear interaction of reliance by frequency,

```
dat.gam2 = gam(AcceptabilityRating ~
               te(RankConstructionVerbReliance, LogFrequencyVerb) + s(RaterGenerosity) +
               s(Verb, bs="re") ,
               data=dat, family=ocat(R=5))
```

and the third model adds `Rater Generosity` to the interaction:

```
dat.gam3 = gam(AcceptabilityRating ~
               te(RankConstructionVerbReliance, LogFrequencyVerb, RaterGenerosity) +
               s(Verb, bs="re") ,
               data=dat, family=ocat(R=5))
```

In these model specifications, `te()` requests a tensor product smooth, a statistical technique for constructing wiggly, nonlinear, prediction surfaces or hypersurfaces (for a non-technical introduction, see Baayen et al., 2017). Below, we show how these complex interactions can be visualized (Figure 1).

As the above three models become gradually more complex, we need to ascertain whether the increase in model complexity is properly counterbalanced by improved model fit. For this, we use the `compareML` function from the `itsadug` package.

```
compareML(dat.gam1, dat.gam2, print.output=FALSE,signif.stars = FALSE)$table
     Model    Score Edf Chisq    Df   p.value
1 dat.gam1 2069.831   8
2 dat.gam2 2063.472   9 6.359 1.000 3.622e-04
compareML(dat.gam2, dat.gam3, print.output=FALSE,signif.stars = FALSE)$table
     Model    Score Edf Chisq    Df   p.value
1 dat.gam2 2063.472   9
2 dat.gam3 2053.897  12 9.575 3.000 2.544e-04
```

The `fREML` scores decrease significantly for both model comparisons, hence the complex three-way interaction of the covariates appears justified. Table 2 presents the summary of this model, but, unfortunately, without providing insight into the nature of the three-way interaction. For this, visualization is essential.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	0.4806	0.0963	4.9904	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(RankConstructionVerbReliance, LogFrequencyVerb, RaterGenerosity)	9.8971	10.7690	219.5628	< 0.0001
s(Verb)	66.8045	93.0000	249.7551	< 0.0001

Table 2: Model summary for the GAMM fitted to the ordered categorical acceptability ratings, with a three-way tensor product smooth.

Figure 1 presents the fitted surface (of predicted acceptability ratings on the scale of the hidden variable u) for the three pairs of two predictors using contour plots. Points with the same expected value are joined by lines. Lower predicted ratings are indicated by deeper shades of grey, whereas higher ratings are visualized with the help of lighter shades of grey. Across the three panels, contour lines are 0.5 rating units apart. The three panels were obtained with the `pvisgam` function from the `itsadug` package.

```
pvisgam(dat.gam3, view=c("RaterGenerosity", "LogFrequencyVerb"),
        select=1, too.far=0.1, main=" ", zlim=c(-2.5, 2.5))
pvisgam(dat.gam3, view=c("RaterGenerosity", "RankConstructionVerbReliance"),
        select=1, too.far=0.1, main=" ", zlim=c(-2.5, 2.5))
pvisgam(dat.gam3, view=c("RankConstructionVerbReliance", "LogFrequencyVerb"),
        select=1, too.far=0.1, main=" ", zlim=c(-2.5, 2.5))
```

Inspection of Figure 1 immediately reveals that the strongest interactions arise with `Rater Generosity`. Ratings increase, unsurprisingly, with `Rater Generosity`, but, as anticipated, this increase is modulated by the frequency and reliance measures. The left panel shows that the effect of `LogFrequencyVerb` changes with increasing `RaterGenerosity`, with a positive slope slowly reversing into a negative slope. Apparently, it is only the less generous raters for whom a greater frequency of use of the verb prompts somewhat higher ratings.

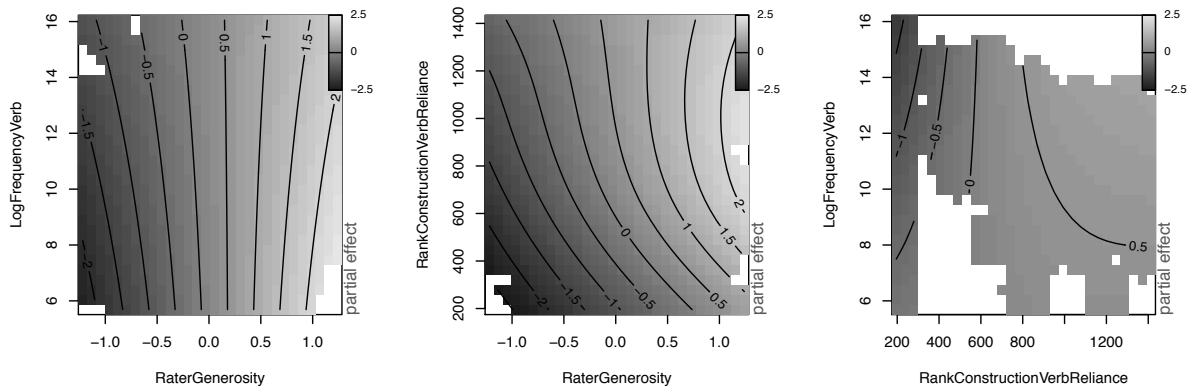

Figure 1: Interaction of Rater Generosity, Log Frequency Verb, and Rank Construction Verb Reliance in the generalized additive mixed model fitted to the acceptability ratings for Polish constructions. Darker shades of grey indicate lower rating scores, lighter shades of grey represent higher rating scores.

The second panel shows a stronger interaction (as evidenced by many more clearly non-parallel or nearly parallel contour lines) showing that the least generous raters are most likely to produce higher acceptability ratings for increasing `RankConstructionVerbReliance`. For the generous raters, we may be seeing a ceiling effect, especially for the larger reliance values; their scores are already so high, that it becomes difficult to up them further.

The third panel of Figure 1 shows the much smaller but nevertheless well-supported interaction of `RankConstructionVerbReliance` by `LogFrequencyVerb`. Reliance has little effect for low-frequency verbs, but gives rise to higher acceptability ratings for constructions with higher-frequency verbs. Furthermore, for verbs with low reliance values, a higher verb frequency leads to reduced acceptability ratings. Apparently, a solid verb frequency without concommittant constructional reliability is an index of unacceptability.

We conclude with addressing the question of whether the model succeeds in predicting the observed ratings. To do so, we extract the model predictions,

```
preds = predict(dat.gam3,dat,type="response",se=TRUE)
head(preds$fit,4)
        [,1]      [,2]      [,3]      [,4]       [,5]
1 0.15171462 0.2283742 0.2763259 0.2176323 0.12595301
2 0.27320177 0.2898621 0.2375499 0.1352225 0.06416371
3 0.09294173 0.1670158 0.2626120 0.2764548 0.20097574
4 0.13646081 0.2149241 0.2765976 0.2317945 0.14022297
```

and as the next step we also extract the ratings with the highest probabilities:

```
maxpos = apply(preds$fit, 1, FUN=function(v)which(v==max(v)))
maxpos[1:4]  # for rows 1 through 4, the index of the column with the highest probability
1 2 3 4
3 2 4 3
```

We tabulate the bin indices with the highest probabilities against the observed ratings

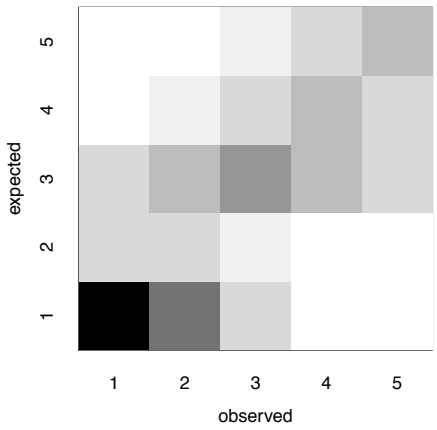

Figure 2: Heatmap for observed versus expected rating classes. Darker shades of grey indicate larger counts.

```
xtabs(~dat$AcceptabilityRating + maxpos)
                        maxpos
dat$AcceptabilityRating   1   2   3   4   5
                      1 205  58  64  10   5
                      2 116  57  82  32   7
                      3  57  40 109  63  28
                      4  25  25  92  73  51
                      5   4  18  54  67  72
```

```
tab = table(dat$AcceptabilityRating, maxpos)
```

and visualize using a heatmap plot (Figure 2). Although most observations are clustered around the diagonal, there is considerable imprecision. Nevertheless, an accuracy of 36.49 is acceptable for a five-way classification problem. It is noteworthy that a GAMM straightforwardly fitted to integer-valued ratings predicts a very similar three-way interaction, but with an overall accuracy that appears to be reduced to around 28.9%.

4 Concluding remarks

The famous statistician George Box is well known for his aphorism "all models are wrong, but some models are useful" (Box, 1976, 1979). Although approaching acceptability ratings with the magnifying glass of the generalized additive mixed model provides us with novel insights, it is advisable to keep in mind that models, by their nature, are simplifications of the true complexity of the data we seek to understand. Although 'ordinal' GAMMs are a tremendous step forward, they provide a window on our data that is still restricted by our current imagination. With this caveat in mind, it is nevertheless exciting to see that the techniques that our colleagues statisticians are developing have so much to offer to our field. The ordinal GAMM laid bare effects of frequency on acceptability ratings that were difficult to model with traditional ordinal regression models (Divjak,

2016). If we follow (Divjak, 2016) and interpret `RaterGenerosity` as a measure of how accepting a subject is with respect to alternative expressions in her language (for further experimental results on 'permissiveness' Dabrowska, 2012; Geeraert, 2016, see also), then the present analysis suggests that less permissive (and more 'prescriptive') subjects are influenced more by frequency of occurrence than more permissive subjects. In the same way, less permissive subjects seem to be more sensitive to Reliance. These findings raise the question of how exposure to language relates to rater generosity. Who is more forgiving when it comes to alternative ways of expression? And what are the consequences of the answer to this question on the perspective we have on the competence of native speakers? Is competence the mastery of a broad spectrum of grammatical variation, or the exclusive mastery of a single idiolect? The present study does not address these questions, but basic demographic information, such as year of birth, native language, parental education, major subject studied, and handedness, are available from the (Divjak, 2016) dataset. The GAM-based technique for probing ordinal acceptability ratings introduced here make it possible for linguists to start exploring these intriguing issues in more depth.

References

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Arppe, A. and Järvikivi, J. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*, 3(2):131–159.

Baayen, R. H. (2011a). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, 11:295–328.

Baayen, R. H. (2011b). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5:436–461.

Baayen, R. H., Vasishth, S., Bates, D., and Kliegl, R. (2017). The cave of shadows. addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Bader, M. and Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46(02):273–330.

Bannard, C. and Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19:241–248.

Bermel, N. and Knittl, L. (2012a). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory*, 8(2):241–275.

Bermel, N. and Knittl, L. (2012b). Morphosyntactic variation and syntactic constructions in czech nominal declension: corpus frequency and native-speaker judgments. *Russian linguistics*, 36(1):91–119.

Bondaruk, A. (2004). *PRO and Control in English, Irish and Polish: A minimalist analysis*. Wydaw: KUL, Lublin.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71:791–799.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in Statistics*, pages 201–236. Academic Press.

Bybee, J. L. and Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2):323–355.

Bybee, J. L. and Hopper, P. J. (2001). *Frequency and the Emergence of Linguistic Structure*, volume 45. John Benjamins Publishing.

Dabrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2(3):219–253.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25:108–127.

Divjak, D. (2008). On (in) frequency and (un) acceptability. In Lewandowska Tomaszcyk, B., editor, *Corpus Linguistics, Computer Tools and Applications — state of the art*, pages 213–233. Peter Lang, Frankfurt.

Divjak, D. (2016). The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in Polish. *Cognitive Science*, DOI:10.1111/cogs.12335:1–29.

Divjak, D. and Gries, S. T. (2012). *Frequency effects in language representation*, volume 2. Walter de Gruyter, Berlin.

Dziwirek, K. (1998). Reduced construction in universal grammar: Evidence from the Polish object control construction. *Natural Language & Linguistic Theory*, 16(1):53–59.

Dziwirek, K. (2000). Why Polish doesn't like infinitives. *Journal of Slavic Linguistics*, pages 57–82.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, 24(02):143–188.

Freyd, M. (1923). The graphic rating scale. *The Journal of Educational Psychology*, 14:83–102.

Funke, F. and Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24(3):310–327.

Geeraert, K. (2016). *Climbing on the bandwagon of idiomatic variation: A multi-methodological approach*. PhD thesis, University of Alberta, Edmonton.

Geeraert, K., Newman, J., and Baayen, R. H. (2017). Idiom flexibility: experimental data and a computational model. *Topics in Cognitive Science*.

Gries, S. T. and Divjak, D. (2012). *Frequency effects in language learning and processing*, volume 1. Walter de Gruyter, Berlin.

Grzegorczykowa, R. (2006). *Wykłady z polskiej składni*. Wydawnictwo naukowe PWN, Warszawa.

Hasher, L. and Zacks, R. T. (1984). Automatic processing of fundamental information. The case of frequency of occurrence. *American Psychologist*, 39:1372–1388.

Hayes, M. H. S. and Patterson, D. G. (1921). Experimental development of the graphic rating scale. *Psychology Bulletin*, 18:98–99.

Janus, D. and Przepiórkowski, A. (2005). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Waliński, J., Kredens, K., and Goźdź-Roszkowski, S., editors, *The proceedings of Practical Applications of Linguistic Corpora*, Frankfurt. Peter Lang.

Keller, F. (2003). A probabilistic parser as a model of global processing difficulty. In Alterman, R. and Kirsh, D., editors, *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 646–651. Cognitive Science Society, Boston, MA.

Kempen, G. and Harbusch, K. (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of german clauses. In Kepser, S. and Reis, M., editors, *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 329–349.

Kempen, G. and Harbusch, K. (2008). Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in german clauses. In Steube, A., editor, *The discourse potential of underspecified structures*, pages 179–192. Walter de Gruyter.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space–time data: A mixed model approach. *Biometrics*, 62(1):109–118.

Lapata, M., McDonald, S., and Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 30–36. Association for Computational Linguistics.

Przepiórkowski, A. and Rosen, A. (2005). Czech and Polish raising/control with or without structure sharing. *Research in Language*, 3:33–66.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition.* Walter de Gruyter, Berlin.

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In Glynn, D. and Fisher, K., editors, *Quantitative methods in cognitive semantics: Corpus-driven approaches*, pages 101–133. Mouton de Gruyter, Berlin.

Schütze, C. T. and Sprouse, J. (2014). Judgment data. In R. J. Podesva, R. J. and Sharma, D., editors, *Research Methods in Linguistics*, pages 27–50. Cambridge University Press, Cambridge.

Sedlmeier, P. and Betsch, T. (2002). *Frequency processing and cognition.* Oxford University Press, Oxford.

Shaoul, C., Westbury, C. F., and Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija*, 46(4):497–537.

Sprouse, J. (2013). Acceptability judgments. In Aronoff, M., editor, *Oxford Bibliographies Online: Linguistics.*

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2016). itsadug: Interpreting time series and autocorrelated data using gamms. R package version 2.2.

Wood, S. N. (2006). *Generalized Additive Models.* Chapman & Hall/CRC, New York.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73:3–36.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association: Theory and Methods*, 111(516):1548–1575.