# Using ordinal logistic regression to analyse self-reported usage of, and attitudes towards swearwords

*Eiselen, Roald, and Van Huyssteen, Gerhard B*
*Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa*
*{roald.eiselen/gerhard.vanhuyssteen}@nwu.ac.za*

## Abstract

Likert-type data is commonly used in many research fields in humanities: from gauging the usability of different user-interface designs, to determining users' likeliness to vote for a particular political party, to evaluation of course materials – to name but a few examples. Despite its prevalence, there is still some disagreement within the statistics community on whether Likert-type scales are true ordinal variables, and by implication whether parametric tests are legitimate to be used in such cases (Endresen & Janda 2017).

In this paper, we explore one parametric statistical test, viz. cumulative odds ordinal logistic regression (OLR), as an analysis method for self-reported data in the humanities. For illustration purposes, our focus is specifically on data of users' self-reported usage of, and attitudes towards swearwords, with the aim of identifying demographic attributes that are predictive of their usage and/or attitudes.

After a brief description of the data we're using, including how the data is being collected, we give a layman's overview of OLR. Since one of our aims is to demonstrate the usability of OLR, we apply our discussion practically to a step-by-step procedure (based on Laerd Statistics 2015) that could be followed easily. We demonstrate the usefulness of the results in reporting on the usage of, and attitude towards two near synonymous Afrikaans swearwords. We show, amongst others, that the odds ratios that are generated as part of the modelling procedure can be used to draw direct conclusions about specific demographic groups.

Keywords: Likert scale, linguistics, offensiveness, ordinal logistic regression, statistical modelling

## 1 Introduction

Over the last several decades, the use of statistical methods in linguistic investigations have become increasingly common, even the norm in many sub-fields of linguistics (Gries 2015). Deciding on which statistical method to use can be a somewhat daunting task, as the nature of the test, as well as the assumptions associated with the statistical test, can limit the types of tests available to a researcher. These factors, of course, also have a direct impact on the types of analysis and interpretation of the results that can be done.

Several types of analysis are commonly used in linguistic analysis, including the use of descriptive statistics, goodness-of-fit tests, monofactorial designs, and linear modelling (see, amongst others, Baayen 2019; Eddington 2015; Gries 2013). However, the use of generalised (i.e., mixed effect) logistic modelling, which take into account multiple predictor (i.e., independent) variables to predict the value of an outcome (i.e., dependent) variable, has been less prevalent. Given the fact that aspects of language production (speak/write) and perception (hear/read), as well as attitudes such as offensiveness of a word, perceived prominence of a word, etc., can be the result of a combination of factors, it is expected that the use of generalised models could be a valuable statistical tool for the analysis and interpretation of linguistic phenomena (Baayen & Linke 2020; Gries 2021). This would however not be applicable to linguistics only, but also more broadly in other fields of digital humanities. With this in mind, we investigate the use of one particular type of generalised logistic model, viz. cumulative odds ordinal logistic regression (OLR).

OLR is a parametric statistical test which describes the relationship between an ordinal outcome variable (i.e., ordered categorical data), and one or more ordinal, categorical or continuous predictor variables. OLR lets you determine which of your predictor variables have a statistically significant effect on an outcome variable, as well as determining how well the OLR model predicts the outcome variable, given a set of predictor variables. In addition to determining variable interaction and prediction, OLR can easily be interpreted as an odds ratio, which provides an

additional interpretation possibility for applying the results of OLR models in real-world contexts (Friendly *et al.* 2015; Harrel 2015).

To investigate the applicability of OLR for linguistic research, we use data collected from the *What The Swearword?!* (WTS) project [1]. One of the aims of this project is to determine offensiveness ratings for Afrikaans swearwords (i.e., any word or expression that could be offensive to some users in some contexts), which could be relevant for content developers, such as authors, publishers, film producers, etc.

The aim of this paper is to demonstrate the usefulness of OLR for this kind of inquiry. For this exploratory study and for illustrative purposes, we determine for only two near-synonymous swearwords, viz. *feeks* and *helleveeg* ('shrew, vixen, harridan'), the relationship between demographic information, and self-reported usage and attitudes ratings. We specifically want to answer the following questions:

- Can OLR be used to predict the usage of, and attitudes towards swearwords?
- Which predictor variables have a statistically significant effect on the usage of, and attitudes towards these two swearwords?
- Are the predictor variables with a statistically significant effect on a particular outcome variable the same for near synonyms?
- Can the interpretation of odds ratios be used to provide practical advice for content developers regarding swearwords?

To answer these questions, we commence with a brief overview of the data that we are using for purposes of this paper, including discussions on our sampling and collection procedures. Section 3 provides an overview of the four assumptions of OLR, as well as the procedure to follow for OLR modelling. This procedure is then illustrated extensively in 4.1, before we also provide more concise ways of presenting results in 4.2. We conclude with a brief discussion of our conclusions, as well as ideas for future work.

## 2 Swearword data

The WTS project website (vloek.co.za) was designed and developed with the main purpose to collect data from users, while experimenting with a variety of surveys, polls, questionnaires, and other data collection tools. Volunteer respondents, recruited through opportunistic and snow-ball sampling (i.e., via social media), have to register as users to participate as (self-selected) respondents. As of 21 August 2021, there are 2,088 registered users on the website, who are all eligible to participate in the surveys.

### 2.1 Demographics

During the registration process, participants provide demographic information, as well as self-reported information on their religious, political and world views. The selection of these questions and their categories is based on similar psycho-sociolinguistic studies (e.g. Beers Fägersten 2007; Dewaele 2016; Janschewitz 2008; Jay 2000, 2020; Van Sterkenburg 2001; Vingerhoets *et al.* 2013) where statistical relationships between one or more of these factors have been correlated with usage of and attitudes to swearwords. The following information, amongst others, is available for all participants in the study (with options for "other" or "don't want to answer" in some cases):

- Age group (three categories; ordinal)
- Sex (four categories; nominal)
- Gender (three categories; nominal) [2]
- Race (five categories; nominal)
- Length group (eight categories; ordinal)
- Highest qualification (12 categories, nominal)
- Income group (eight categories; ordinal)
- Religious view (five-point scale, from Not religious at all, to Very religious)
- Political view (five-point scale, from Very liberal, to Very conservative)
- World view (pertaining to moral and social issues; five-point scale, from Very liberal, to Very conservative)

Due to our sampling method and mode (social media, and a website), we assumed a priori that our sampling population will not be representative of the general Afrikaans population, since there are some inherent assumptions about this population. These include that they:

- have regular access to a computer/mobile device, and an internet connection;
- are technologically savvy (e.g., they are using social media platforms);
- have an interest in language, and specifically swearword or other taboos;
- are therefore probably less easily offended by such words and taboos (even though they may not use and/or approve of such words); and
- thus perceive themselves as rather enlightened/liberal.

These assumptions are confirmed when we look at the descriptive statistics of the groups that responded to the questionnaires for the two words under consideration (for *feeks* n=133; for *helleveeg* n=90). Only a small percentage of the respondents are 60 or older (21.1% for *feeks* and 18.9% for *helleveeg*); for both questions the entire population is white, and there are more males than females (unlike in the general Afrikaans population [3]); and the population is highly educated (64.7% of the respondents for *feeks* and 73.3% for *helleveeg* have a university degree). Although the entire population for both questions is mostly religious to some degree, only 7.5% (*feeks*) and 8.8% (*helleveeg*) of the respondents identify as conservative or very conservative.

When interpreting any of the results in this project, one should therefore be aware of the fact that the sample population is not representative of the Afrikaans community. Such results should therefore be preferably seen as individual pieces of empirical evidence that should be corroborated with other evidence, to get the full picture of a bigger puzzle.

### 2.2 Collection of self-reported data

One of the project's main types of short surveys, is the single word survey (SWS), where only one swearword per survey is presented to registered participants. The aim with SWSs is to keep each one as short as possible, in order to prevent respondent fatigue – "a well-documented phenomenon that occurs when survey participants become tired of the survey task and the quality of the data they provide begins to deteriorate" (Lavrakas 2008). The assumption is that one would cover more words over a period of time, than if one were to present the same number of words to participants in a single session.

Participants are therefore not required to complete questionnaires on all words, but only those ones that they want to participate in, and/or that they have time for. The implication of this way of sampling is that we cannot assume that (a) the data per word is independent (because some of the respondents might have answered all the SWSs); or (b) the responses are from the same sampling group (because some of the respondents might not have answered all the SWSs). Evaluating the pros and cons of this sampling method is, however, not the focus of this paper, but will be addressed in future research.

To make it as easy as possible for participants, they must self-report their usage of, and attitudes towards a given word on Likert-type scales, which are typically used to collect qualitative data in a way that provides quantitative values, thereby making statistical analysis of the data possible (Dubois 2013). For this study, a 9-point scale was used, where only the scores at the two extreme ends are descriptively categorised; this reinforces the notion that there are equal distances between each point on the scale (Endresen & Janda 2017). Respondents are asked to report their judgments on each of the following eight questions:

1. How often do you *say* or *write* the word? (Never … Very often)
2. How often do you *hear* or *read* the word? (Never … Very often)
3. How *offensive* do you find the word personally? (Not at all … Very)
4. How *taboo* or socially unacceptable is the word for people in general? (Not at all … Very)

5. To what extent is the word *emotionally charged* for you? (Very negative … Very positive)
6. How *prominent* is the word? (Not at all … Very)
7. How well do you know what the word *means*? (Not at all … Very well)
8. Is the word used pertaining to *men only*, *men and women*, or *women only*? (Men only … Women only)

The responses to each of these questions are considered as the outcome variables, while the demographic data are considered as the predictor variables. The hypothesis is that one or more demographic factors (such as age, or political view) will have a statistical effect on the usage of, or attitudes towards the swearwords (see Beers Fägersten 2007; Dewaele 2016; Janschewitz 2008; Jay 2000, 2020; Van Sterkenburg 2001; Vingerhoets *et al*. 2013).

## 3 OLR modelling

OLR modelling is a parametric statistical test to determine whether one or more predictor variables have a statistically significant effect on an outcome variable, and how well the model can predict the value of the outcome variable, given a set of predictor variables (Friendly *et al*. 2015; Harrel 2015; Laerd Statistics 2015). OLR has four assumptions that need to be considered in order to determine if it is a valid statistical approach for a particular study.

The first two assumptions are related to the design of the study and the measurements taken. *Assumption one* requires that you have a single ordinal outcome variable. *Assumption two* states that you should have one or more predictor variable(s) that are continuous, categorical, or ordinal. It should be noted that ordinal predictor variables are treated as categorical (i.e., they lose any internal ordering distinctions as part of the modelling procedure).

The last two assumptions relate to how the data fits the OLR model to provide valid test results. *Assumption three* states that there should be no multicollinearity between two or more continuous predictor variables. This means that if two continuous predictor variables are highly correlated, the results cannot be interpreted accurately, since it will not be possible to determine which one of the two predictor variables contribute to the explanation of the outcome variable. *Assumption four*, which is the fundamental assumption of OLR, states that you must have proportional odds, which means that each predictor variable has an identical effect at each cumulative split in the ordinal outcome variable.

Informed by the procedure suggested by Laerd Statistics (2015), the first step of the OLR modelling procedure is to ensure that the data adheres to the assumptions of the test. The first assumption requires an outcome variable that is ordinal. Although parametric tests, such as OLR, have been applied widely to Likert-type data in various other studies (e.g. Zhou *et al*. 2009), there is some disagreement within the community on whether Likert-type scales are true ordinal variables, and by implication whether parametric tests such as OLR are legitimate to use in such cases (Endresen & Janda 2017). However, Endresen & Janda (2017) show that for Likert-type data, the results for parametric and non-parametric tests have comparable results. With this in mind, we assume that Likert-type data is indeed ordinal, and that this type of parametric analysis is valid. Adherence to the second and third assumption is more easily confirmed, since all the predictor variables (i.e., the demographic information) are ordinal or categorical.

Verifying adherence to the fourth assumption is relatively easily tested in a statistical package such as SPSS by using "Test for parallel lines". This test (also known as a full likelihood ratio test) compares the fit of the proportional odds model to a cumulative odds model without the proportional odds assumption. If the assumption is met, the Chi-square value of the model will be small and not statistically significant ($p > 0.05$). Any variables that violate this assumption must be excluded from an OLR model.

After removing all predictor variables that violate any of the assumptions, the OLR is run, using an appropriate statistical package (SPSS in our case). The OLR test produces three important test results that should be reviewed before

investigating the full set of model parameter estimates:

1. a deviance goodness-of-fit test, which indicates if the model is a good fit for the data, where larger values are more indicative of a good fit;
2. an omnibus test (the likelihood ratio test [4]), which indicates whether the model predicts the outcome variable statistically significantly better than an intercept-only model (i.e. a model that does not take predictor variables into account); and
3. the effects of the different predictor variables, by looking at the Wald $\chi^2$ test statistic and associated statistical significance (where $p < 0.05$).

Next, depending on the statistical significance of the model fit, and the effect of the different predictor variables in the model, additional predictor variables that clearly do not have an effect on the outcome variable, could be removed – both to simplify the model, and to improve the fit of the model. Therefore, given the results, one can either report the model as is, or try to improve the model by only selecting a subset of the predictor variables to see if there is any improvement in the overall fit of the model. However, care should be taken, since there are often intervariable effects, which might mean that a combination of predictor variables (e.g. gender plus age) could create a better model fit, even though one of these predictor variables does not have a statistically significant effect on the outcome variable.

The final step in the procedure is to interpret the predictor variable parameter estimates for each category of the predictor variables, and their significance. This interpretation should provide insight into the specific effect of each category of that predictor variable on the outcome variable.

## 4 Examples

### 4.1 Extensive example of OLR procedure

For the purposes of illustrating the procedure described in the previous section, we select one of the words, *feeks*, and one outcome variable, Tabooness ("How taboo or socially unacceptable is the word for people in general?"), as an application example for the full procedure. Additional, more concise examples of results are presented in section 4.2.

*Table 1: Test of model effects: Tabooness of "feeks"*

| Predictor variable | *Wald* $\chi^2$ | *df* | *Significance* |
|---|---|---|---|
| Age | 3.340 | 2 | .188 |
| Gender | 11.153 | 1 | .001 |
| Length | 11.507 | 5 | .042 |
| Country | 14.961 | 8 | .060 |
| Political view | 11.101 | 4 | .025 |

*Step 1: Determine if the data adheres to the assumptions of the OLR test*

Given that the outcome variable is ordinal (i.e., data on a 9-point Likert scale), and all predictor variables are categorical, the first three assumptions of OLR are adhered to. For the fourth assumption, all predictor variables are tested for violation of the proportional odds assumption. For the word *feeks* and the Tabooness outcome variable, four of the predictor variables violate the assumption of proportional odds, viz. Qualification, Income, Religious view, and World view. Five variables do not violate this assumption, and will therefore remain in the initial OLR model.

*Step 2: Run OLR and review results*

For the Tabooness outcome variable and five predictor variables, the deviance goodness-of-fit test indicated a good fit of the observed data $\chi^2(748)=408.662$, $p=.546$, and the likelihood ratio test does statistically significantly predict the outcome variable over and above the intercept-only model, $\chi^2(20)=39.284$, $p=.006$. The model effects produced by OLR, presented in Table 1, show that three of the variables have a statistically significant effect on the outcome variable, Gender ($p=.001$), Length ($p=.042$) and Political view ($p=.025$). Age ($p=.188$) and Country ($p=.060$) do not show statistical significant effect on the outcome variable, although Country does account for the most data.

*Step 3 (optional): Exclude uncorrelated predictor variables to simplify the model, and improve its fit*

*Table 2: Parameter estimates table (condensed): Tabooness of "feeks"*

| *Parameter* | *Beta* | *Wald* $\chi^2$ | *Sign.* | *Exp(B)* | *Odds* |
|---|---|---|---|---|---|
| Gender: Male | -1.209 | 11.191 | .001 | .299 | 1:3.34 |
| Gender: Female | 0 | . | . | 1 | 1:1 |
| | | | | | |
| Political: Very conservative | -2.891 | 4.337 | .037 | .056 | 1:17.85 |
| Political: Very liberal | -1.346 | 8.769 | .003 | .260 | 1:3.85 |
| Political: Moderate | -.461 | 1.359 | .244 | .631 | 1:1.58 |
| Political: Conservative | .446 | .408 | .523 | 1.562 | 1.56:1 |
| Political: Liberal | 0 | . | . | 1 | 1:1 |

In the case of Tabooness of *feeks*, we removed the variables that do not show statistically significant effect (Age and Country), but this decreased the fit of the model ($p$=.010). Therefore, keeping all five variables produced the most statistical significant fit for the observed data. This is most likely due to the fact that there are interactions between predictor variable groups, e.g. including both Age and Gender, that contribute to the overall model fit.

*Step 4: Interpret results by reviewing parameter estimates, and determining the odds-ratios for values of specific predictor variables*

Table 2 provides a condensed view of the most important information from the parameter estimates table for two variable groups, viz. Gender and Political view. The table includes the beta, Wald $\chi^2$, significance, and odds-ratio (Exp(B)) values.

Since we have established that the OLR model for Tabooness of *feeks* fits the observed data, we can now interpret the information in the parameter estimates table for effects between specific groups of respondents and the outcome variable.

Keep in mind that OLR expresses parameter estimates in terms of one reference group (i.e., one of the categories under a variable). For each predictor variable, one category is selected as the reference group, and no beta or significance values are calculated for such a selected category; in Table 2 these are Gender: Female, and Political: Liberal. The Exp(B) value represents the odds ratio, i.e., the odds that that group will either assign a higher score (values larger than 1), or a lower score (values smaller than 1). As an example: For the word *feeks*, the odds that a man will assign a lower Likert score than a woman, are 3.34 (1/.299) times, which is a statistically significant effect, $\chi^2(1)=11.191$, $p$=.001. In other words, we could expect that women are more likely to rate *feeks* with a higher taboo score than men.

Another example: The odds of people who are Very liberal to assign a lower Likert score than people who are Liberal, are 3.85 times, also a statistically significant effect, $\chi^2(1)=8.769$, $p$=.003. In contrast, politically conservative respondents are 1.56 time more likely to assign a higher score than liberal respondents, but this is not a statistically significant effect ($p$=.523).

In the following section we apply the same procedure to two outcome variables for both *feeks* and *helleveeg*, to show how results can be more concisely reported. We also illustrate further interpretations of the results.

### 4.2 Concise examples of results

Given two words, *feeks* and *helleveeg*, and eight outcome variables, a total of 16 OLR models are possible. Since the aim of this paper is to demonstrate the applicability of OLR models to this type of inquiry, and for the sake of brevity, we report on the OLR tests and procedures for only two outcome variables, namely:

- How often do you *hear* or *read* the word? (Hear/Read)
- How *prominent* is the word? (Prominence)

As discussed in the previous section, the first three assumptions of OLR are not violated, since the outcome variables are all Likert-type data, and all

the predictor variables are categorical in nature. For the fourth assumption, all variables are tested for violation of the proportional odds assumption. For both outcome variables, across both words, many of the variables violated the assumption of proportional odds, and therefore cannot be included in the remainder of the procedure. A summary of the remaining predictor variables for each outcome variable for both words is provided in Table 3.

*Table 3: Remaining predictor variables after testing for assumption of proportional odds*

| | *Feeks (n=133)* | *Helleveeg (n=90)* |
|---|---|---|
| Hear/Read | Age<br>Gender<br>Length<br>Country<br>Qualification<br>Religious view<br>Political view | - |
| Prominence | Income<br>Religious view<br>Political view<br>World view | Age<br>Gender<br>Length<br>Country<br>Income<br>Political view |

The first thing to note from this table, is that there is no overarching set of predictor variables that adhere to the proportional odds assumption for both words across the two outcome variables. Separate models and variable selection are therefore necessary for each swearword, and for each outcome variable.

Also note that all the predictor variables for *helleveeg* violate the proportional odds assumption for the Hear/Read outcome variable. This stems from the fact that the distribution of assigned scores is very skewed, and 74.4% assigned either a 1 or 2 on the scale, indicating that they never or very rarely read or hear the word. [5] For *feeks,* on the other hand, there is a much more equal distribution across the various scale scores, with between 9% and 15.3% of responses in 7 of the 9 scale scores. Although there is no inherent assumption about the distribution of data for OLR, in cases where the distribution is highly skewed on the outcome variable, it is likely that either all the predictor variables will violate the proportional odds assumption, or that the resultant model will not be significantly better than an intercept-only model.

Given these remaining predictor variables, we firstly create OLR models that include all of the predictor variables that are valid for the OLR test. We then review the first three statistical tests to determine (a) the fit; (b) whether the model performs statistically significantly better than an intercept-only model; and (c) what the effects of the different predictor variables on the outcome variables are.

The following subsections provide the results for the words *feeks* and *helleveeg* for the two outcome variables, where only the best model for each outcome variable is described and interpreted. The aim is to illustrate that the entire statistical procedure can be expressed much more succinctly for each set of outcome and predictor variables.

*Feeks*

An OLR was run to determine the effect of Length, Qualification, Religious view, and Political view on how often participants Hear/Read the word *feeks*. There were proportional odds as assessed by a full likelihood ratio test comparing the model with varying location parameters, $\chi^2(154)=175.326$, $p=.115$. Although the deviance goodness-of-fit test indicated that the model was a good fit of the observed data, $\chi^2(866)=478.641$, $p=1.00$, the final model did not statistically significantly predict the outcome variable over and above the intercept-only model, most likely due to the high rate of empty cells for combinations of predictor variables (> 50%) [6]. Various models with fewer variables, which decrease the empty cell rate, also did not improve the fit of the overall model significantly.

For the Prominence of *feeks*, an OLR was run to determine the effects of Religious view [7]. The full likelihood ratio test indicated that there were proportional odds, $\chi^2(28)=17.670$, $p=.934$, while the deviance goodness-of-fit test also indicated that the model was a good fit of the observed data,

$\chi^2(28)=18.601$, $p=.664$. The overall model statistically significantly predicted the outcome variable over and above the intercept-only model, $\chi^2(4)=18.049$, $p=.001$. The odds of respondents scoring prominence lower than Religious respondents are statistically significant for two categories: Average religious, 2.81 times, ($p=.030$), and Not at all religious, 4.30 times ($p=.001$). The odds that Very religious participants would rate *feeks* higher on the Likert scale, is 1.41 times, but it is not statistically significant ($p=.425$). The results from this model indicate that more religious people are more likely to find the word *feeks* prominent when compared to people who are less religious.

*Helleveeg*

Since no variable adhered to the assumption of proportional odds for the Hear/Read variable, an OLR was only run for Prominence to determine the effects of the variables listed in Table 3. The first model, which included all six variables, did not predict the outcome variable statistically significantly over and above the intercept-only model, $\chi^2(26)=38.909$ and $p=.05$. By removing the predictor variable with the least effect, Gender, the model did improve, $\chi^2(25)=38.895$, $p=.038$, and statistically significantly predicted the outcome variable over and above the intercept-only model. Of the predictor variables, Age accounted for the largest proportion of the data: respondents between the ages of 40 and 59 were 4.54 times more likely to find *helleveeg* prominent, than people over the age of 60, a statistically significant effect ($p=.019$). Although the odds of people under the age of 40 is 1.47 times more likely to find the word more prominent, this effect is not statistically significant ($p=.539$).

## 5 Conclusion

In this paper, we have demonstrated that OLR is able to generate models that, in some instances, statistically significantly predict the outcome variable over and above an intercept-only model. Using data from three questions for two words, OLR was able to identify demographic variables that have a statistically significant effect on the Likert scale for the three questions. However, the demographic variables that have a statistically significant effect varies for both the different questions, and the different words. This is partly due to the fact that different predictor variables violate the primary assumption of proportional odds and therefore cannot be included in the OLR model. We concluded that it is essential to do rigorous testing of adherence to OLR's four assumptions for all predictor variables, in order to ensure that the OLR model is valid.

Beyond the differences in the predictor variables for the different questions and words, we also found that different numbers of variables are required to find the best fit for the data. In some cases, such as the prominence of *feeks*, a single predictor variable created the best model, while five variables were necessary for the Tabooness model of *feeks*.

Given the fact that the same variables do not have significant effects for the different words, we concluded that our kind of data and sampling methods do not allow to directly compare the OLR results or models of different words – at least at this stage of our research. This could possibly be due to two aspects:

1. Data for the two words were collected from two different, but potentially overlapping sampling groups. We expect intuitively that there should be larger overlaps of predictor variables (e.g. we might expect that very religious people will rate most swearwords more offensive than people who are perhaps less religious).
2. The semantic fields of different swearwords might also play a role. Near synonymous swearwords come from the same semantic domain (e.g. RELIGION) and we might expect that their tabooness ratings will all depend on similar predictor variables.

Since we have not observed these expectations in the results above, we will need to investigate how to deal with these anomalies in future studies. Other or additional statistical tests will most probably be needed to allow direct comparisons between the outcome variables for different words (see Van Huyssteen & Eiselen, 2021).

Based on the OLR models that have been created for the respective questions and words, we have shown that the odds ratios that have been generated as part of the modelling procedure, can be used to draw direct conclusions about specific demographic groups. For example, what are the odds that women will rate *feeks* as more Taboo than men, or that people under the age of 60 will find *helleveeg* more prominent than older respondents.

Although these are encouraging results for using OLR to investigate the usage of, and attitudes towards swearwords, several outstanding issues need to be addressed to determine how well this type of modelling works for this kind of data. In addition to matters already mentioned above, these include:

- the applicability of OLR modelling to other swearwords that are used more often and are more well-known;
- how sample size and demographic distribution affect the model's descriptive quality;
- how data distribution affects the ability of the models to identify variable effects;
- model visualisations that make the data and results more accessible to publishers and writers; and
- whether the models will be more or less useful indicators of variable effect on smaller Likert scales, such as a 3- or 5 point scale.

## Notes

[1] A comprehensive overview of this project is provided in another paper (submitted for presentation) at this conference. See Van Huyssteen, 2021.

[2] The question is: "Do you identify with one or more specific gender groups?", with options Yes, No, Don't want to answer. If a respondent choose Yes, they can specify which group(s).

[3] The ratio male:female for both words was 57:43. For the general South African population, the ratio in the 2011 Census was 49:51. Based on data in Centre for Risk Analysis (2020), we can calculate that of the total white population in South Africa, 29.8% males and 31.5% females consider Afrikaans their first language.

[4] Two separate likelihood tests are performed as part of the OLR procedure, and they should not be confused with one another. The first, referred to as the full likelihood ratio test, is an assumption test for proportional odds; the second determines the fit of the full model.

[5] This is corroborated by data from all the corpora on VivA-KPO (2021): the distribution *feeks*:*helleveeg* is 93:7 per hundred examples.

[6] Empty cells in this context refers to a combination of predictor variables with no respondents, e.g. a person who is taller than 199cm (Length), has a doctorate (Qualification), is very conservative (Religious), and is very liberal (Political view). High rates of these empty cells, about which no statistical information is available, can be detrimental to the quality of the model and usually occurs if the sample group is relatively small, and a large number of variables, with a large number of categories are included in the model.

[7] The categories for Religious view are: Not religious at all; Not particularly religious; Average religious; Religious; Very religious. Respondents also had the option to specify something else, or to choose not to answer the question.

## Acknowledgements

This research is partially funded by the Suid-Afrikaans Akademie vir Wetenskap en Kuns, and partially made possible through barter agreements with BlueTek Computers, Afrikaans.com, and WatKykJy.co.za.

Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the North-West University (ethics number: NWU-00632-19-A7).

The second author is a director of the not-for-profit company Viridevert NPC (CIPC registration number: 2016/411799/08), who owns and manages the website vloek.co.za. This website was developed specifically for this project, and this conflict of interest has been approved by North-West University.

We would like to thank Jaco du Toit (NWU) for his help with data processing. Thank you also to Adam Lund (Laerd Statistics) for suggesting OLR for our kind of data.

None of the results and/or opinions in this paper can be ascribed to any of the people or organisations mentioned above.

## References

Baayen, RH 2008, *Analyzing linguistic data: A practical introduction to statistics using* R, Cambridge University Press, Cambridge

Baayen, RH & Linke, M 2020, 'Generalized Additive Mixed Models', in Paquot, M & Gries, ST (eds.) *A Practical Handbook of Corpus Linguistics*, Springer Nature, Cham, pp. 563-591.

Beers Fägersten, K 2007, A sociolinguistic analysis of swearword offensiveness, Universität des Saarlands, Saarbrücken, view 16 August 2021, <https://www.researchgate.net/publication/265009714_A_sociolinguistic_analysis_of_swearword_offensiveness>.

Centre for Risk Analysis 2020, *Socio-Economic Survey of South Africa*, Melville, view 16 August 2021, <https://cra-sa.com/products/socio-economic-survey/2020>.

Dewaele, J-M 2016, 'Self-reported frequency of swearing in English: do situational, psychological and sociobiographical variables have similar effects on first and foreign language users?', *Journal of Multilingual and Multicultural Development*, vol. 38, no. 4, pp. 330-345.

Dubois, D 2013, 'Statistical reasoning with set-valued information: Ontic vs. epistemic views', in Borgelt, C, Gil, MA, Sousa, JMC, & Verleysen, M (eds.) *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, Springer, Berlin, pp. 119–136.

Eddington, D 2015, *Statistics for linguists: a step-by-step guide for novices*, Cambridge Scholars Publishing, Newcastle upon Tyne.

Endresen, A & Janda, LA 2017, 'Five statistical models for Likert-type experimental data on acceptability judgments', *Journal of Research Design and Statistics in Linguistics and Communication Science,* vol. 3, no. 2, pp. 217–250. https://doi.org/10.1558/jrds.30822.

Friendly, M, Meyer, D & Zeileis, A 2015, *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*, 1st edn, Chapman and Hall, Boca Raton.

Gries, ST 2013, *Statistics for Linguistics with R: A Practical Introduction*, 2nd edn, De Gruyter, Berlin.

Gries, ST 2015, 'Quantitative linguistics', in Wright, J. (ed.), *International Encyclopedia of the Social and Behavioral Sciences*, 2nd edn, Vol. 19, Elsevier, Oxford, pp. 725–732.

Gries, ST 2021, '(Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example', *Language Learning*, vol. 71, no. 3, pp. 757-798.

Harrel, FE 2015, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression and survival analysis*, 2nd edn, Springer, Heidelberg.

Janschewitz, K 2008, 'Taboo, emotionally valenced, and emotionally neutral word norms', *Behavior Research Methods*, vol. 40, no. 4, pp. 1065-74.

Jay, T 2000, *Why we curse: A neuro-psycho-social theory of speech*, John Benjamins, Amsterdam.

Jay, T 2020, 'Ten issues facing taboo word scholars', in Nassenstein, N and Storch, A (eds.) *Swearing and Cursing*, De Gruyter Mouton, Berlin, pp. 37-52.

Laerd Statistics 2015, 'Ordinal logistic regression using SPSS Statistics', *Statistical tutorials and software guides*, view 16 August 2021, <https://statistics.laerd.com/>.

Lavrakas, PJ 2008, *Encyclopedia of survey research methods* (Vols. 1-0), Sage Publications, Thousand Oaks, doi: 10.4135/9781412963947.

Van Sterkenburg, PGJ 2001, *Vloeken. Een cultuurbepaalde reactie op woede, irritatie en frustratie*, 2nd edn, Sdu Uitgevers, The Hague.

Vingerhoets, AJJM, Bylsma, LM & De Vlam, C 2013, 'Swearing: A Biopsychosocial Perspective', *Psychological Topics*, vol. 22, no. 2, pp. 287-304.

VivA-KPO 2021, Virtual Institute for Afrikaans: Corpus Portal Comprehensive, view 19 August 2021, <http://viva-afrikaans.org>.

Van Huyssteen, GB 2021, 'Swearing in South Africa: Multidisciplinary research and scientific communication on language taboos', *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*, 29 November to 3 December, DHASA, South Africa.

Van Huyssteen, GB, Eiselen, ER 2021, (In print), 'Oor feekse en helleveë [On shrews and harridans]', *Tydskrif vir Geesteswetenskappe*.

Zhou, F, Wu, D, Yang, X & Jiao, J 2008, 'Ordinal logistic regression for affective product design', IEEE International Conference on Industrial Engineering and Engineering Management, IEEE, pp. 1986-1990.