Bollestad, Ole Alexander Svihus

Using GLM to Analyze Acceptability Judgement Test Data

A case study of The role of L1 Norwegian and L2 English in the acquisition of verb placement in L3 German.

NTNU
Norwegian University of
Science and Technology

Bollestad, Ole Alexander Svihus

Using GLM to Analyze Acceptability Judgement Test Data

A case study of The role of L1 Norwegian and L2 English in the acquisition of verb placement in L3 German.

NTNU
Norwegian University of
Science and Technology

Using GLM to Analyze Acceptability Judgement Test Data

A case study of The role of L1 Norwegian and L2 English in the acquisition of verb placement in L3 German.

Bollestad, Ole Alexander Svihus

08.02.2020

Abstract

In this thesis we have evaluated linguistic research question using GLM models and proportion correct answers of sentences from Acceptability Judgement Test data done by upper secondary school students with one,two, four and five years learning German. The aim is to see if GLM and offer more interpretation opportunites for linguistic research and to analyze the data on this form. To evaluate the research questions we have used representations of the distribution to find evidence of transfer from English sentence structure, and using the GLM models to determine whether the proportion correct answers for German sentences with topicalization or sentence adverbial as response variable is influenced by the proportion correct answer for English or other German sentences as explanatory variables. We found that as a statistical method it gives new opportunities for interpretation.

Acknowledgements

I would like to thank my supervisor Ingelin S. for her dedicated support and guidance. Ingelin continuously provided encouragement and was always willing and enthusiastic to assist in any way she could throughout this thesis. I would also like to thank Anne D. for providing the data and her research that has been used in this thesis. Finally, many thanks for all the positive messages from friends and family through my work.

Contents

1 Introduction

How well do we learn languages given high proficiency in a native and secondary language? This is a highly discussed topic in linguistic research, and there are several research approaches to it. Being a fairly new topic of interest, there has not been a general consensus of how to study it (Bardel & Falk, 2007). Bardel and Falk (2007) mentions some papers that have studied similarities between languages in the same family and used the shared vocabulary, while others that have looked at the vocal communication. Acquisition of language through same family of languages is called Universal Grammar (UG), and means that languages in the same family often have the same grammar structure (Bardel & Falk, 2007). Lexical learning is another, and is the learning through understanding of words and how it is used in the language of fluency. Bardel and Falk (2007) also brings up the theory of Interlanguage Grammer (ILG). This is the learners construction of language from transfer, and is especially noticable for learners fluent in a language outside the given language family. The focus in this thesis has been on the shared language between Norwegian (L1) and German (L3), with the main focus being on transfer from assumed high profiency in second language English (L2). We are looking if there is evidence of L2 transfer when learning L3. For the most part we assume transfer through UG, but can not rule out lexical or ILG.

This paper is based on a study by Dahl (2020) on the role of L1 and L2 speakers acquisition of verb placement in L3. We aim to look for transfer of already learned languages under acquisition of a new language. As a language, all three are from the germanic family, but the verb placement is different between L2 and L3 in sentences. In both L1 and L3 we typically see verbs placed in the second position (V2), while for L2 verb placement is in third position (V3). Thus we say that the transfer from L1 would be target-like (Dahl et al., 2020). Participants in the study are high school students with varying experience with L3 speaking, having finished one, two, four and five years of

German. Studies show that spending time using the language increases the ability in acquisition. Also that the general knowledge about learning a language makes it faster and better to learn (Jackson & Kaplan, 1999). With this in mind we would expect the participants with more years learning L3 to do better.

The study in focus has used an acceptability judgement test to gather the data. An acceptability judgement test (AJT) is a test where you ask participants to judge whether a sentence is acceptable or not. To be acceptable in this case means that the participant is showing signs of competence speaking the language by judging something as (un)grammatically correct. This kind of test is claimed by many researchers to inform us more about the linguistic competence in the target language for the participants. In the study a Likert scale has been used for the participant to judge with. This is a summated rating approach with a response categorization system of up to 7 alternatives typically. In our data a scale 1 to 4 is used. Likert is described as an enjoyable kind of scale,which in many situations can ensure internal consistency and ability to differentiate between individuals (Robson & McCartan, 2016). This scale is also benefitiary for statistical methods like *analysis of variance*, or ANOVA, which is the most used method to analyse linguistic data. We want to extpand these methods in this thesis and study the potential benefits of using Generalized Linear Models (GLM). To do this we introduce a transformation of the Likert scale data into proportion correct answers for each group of sentences in the data, described in detail in Chapter 3.1.1.

The type of sentences we are analysing German sentences with topicalization and sentence adverbial (see Chapter 3.1.1), which participants judged for grammatical and ungrammatical sentences in the AJT. We have used a three-trial binomial probability approach to find proportion correct answers, described in Chapter 2. This is done by assuming a Likert score as correct, and adding the grammatical and ungrammatical correct judgements together estimating the total proportion of correct answers in the test.

We further on analyze these proportions focusing on presenting the distribution in plots (Ch. 3.2) and using them as variables to estimate the proportion in Chapter **??**.

1.1 Working hypothesis

In this thesis we want to explore the use of GLM as a potential statistical tool for analysing linguistic data. To do this we got the working hypothesis that;

- *GLM and proportion of correct answers can be used to analyze AJT-data, and give new interpretation opportunities compared to methods traditionally used.*

We base the working hypothesis on exploring different research questions, shared by Dahl (2020) in her research.

RQ1 Is there evidence of L1 and/or L2 transfer to L3 in Norwegian L3 learners of German?

RQ2 If so, is this transfer equal in sentences with topicalizations and in sentences with sentence adverbials?

RQ3 Does mastery of the relevant structure in L2 matter for whether it is eligible for transfer?

RQ4 Does L2 proficiency influence transfer in L3; and if so, does higher proficiency predict more or less transfer from L2?

RQ5 Does evidence of L2 transfer become more or less pronounced with increasing L3 proficiency?

2   Analysis Method

As the data is on a Likert score format, we want to transform this data into proportions before using them in models and to analyze focusing on the research questions. For a lot of the judgements we notice an almost randomised distribution when we expect the participants to be distributed among the desired judgement (see Chapter 3.2). To reduce some of this randomness we are using a three-trial method to model the probabilities. We assume that every participant gets the same sentence 3 times during the AJT, and thus have a chance to get it right 3 times. From the data we put a score for the judgement they have already done for the sentence. If a grammatical sentence is judged as 4, we assume that they will get this sentence correct 3 of 3 times, and give this participant a score of 3. If a participant judge a grammatical sentence as 3 they get a score of 2, as we assume they will judge it correct 2 times. A Likert score of 1 for a grammatical sentence will give this participant a score of 0 from this method.

For ungrammatical sentences this scale is turned around. Meaning that if a participant judge an ungrammatical sentence as 1, they will be get a score of 3. In our data of proportions the scores from both are added in a total proportion.

Doing this we increase our dataset by 3 times, and the way we find the probabilities changes from a strict "Success" and "Failure" binomial model to a model with multiple trials. We get the binomial probability function,

$$Bin(r_i, \pi_i, 3) = \binom{3 \times n}{r_i \times n} \pi_i^{r_i} (1 - \pi_i)^{3 - r_i} \tag{1}$$

where the r_i are the individual scores in the interval $[0, 3]$ every participant gets from their initial judgement, n is the number of sentences and π_i are the estimated probability of "Success" for every individual. In our models $n = 12$ for every proportion.

2.1 RQ1 and RQ2

The research questions in focus for this section are; RQ1: *Is there evidence of L1 and/or L2 transfer to L3 in Norwegian L3 learners of German?* and RQ2: *If so, is this transfer equal in sentences with topicalizations and in sentences with sentence adverbials?*

The basis of this study is that L1 Norwegian and L3 German share approximately the same sentence structure when it comes to placement of verbs, while L2 English is different (Dahl et al., 2020). RQ1 are outside the scope of this thesis to do formal evaluation, but there are indications in Chapter 3.2, and will be discussed there. A negative relationship between English and German sentences may indicate that there is transfer between the two. We are restricting the analysis to sentences with topicalizations and sentence adverbials, as is the focus in RQ2.

2.2 RQ3

For this section the research question in focus is; *Does mastery of the relevant structure in L2 matter for whether it is eligible for transfer?*

To determine whetcher a participant is showing mastery of the relevant structure in L2, we can look at the proportion correct judged data for the English sentences. A high proportion correct may indicate that a participant is confident in his or her own ability to judge a sentence correct, showing mastery of the structure. If mastery of the structure makes it eligible, then there will be participants that are judging more German sentences incorrect when they have high proportion of English sentences correct. Dahl (2020) points out that we expect to see more transfers of V3 for sentences with topicalizations than for sentences with sentence adverbials, which correlates this research question with RQ2, RQ4 and RQ5.

2.3   RQ4 and RQ5

For this section we want to look at the research questions; RQ4: *Does L2 proficiency influence transfer in L3; and if so, does higher proficiency predict more or less transfer from L2?* and RQ5: *Does evidence of L2 transfer become more or less pronounced with increasing L3 proficiency?*

These research questions both aim to investigate any correlation between being proficient in L2 and if this predicts more or less transfer. They are the main focus of this thesis, and will be investigated further in Chapter 5 using GLM (Chapter 4.1) to create models that can predict the proportion correct judgements based on explanatory variables involving the proportion correct judged English sentences. For RQ5 we increase the models to include proportion correct German sentences as well. RQ5 is evaluated thoroughly by adding an interaction term in the models, which hopefully help us observe an effect between English and German sentences.

3   Case study: The role of L1 Norwegian and L2 English in the acquisition of verb placement in L3 German

3.1   AJT dataset

3.1.1   Intro

In this research paper the data considered were provided by the work done by Dahl (2020) in an unpublished draft. This was done giving participants AJT for English and German sentences. The participants were 154 male and female students between 16-17 years attending upper secondary school in Norway. In the dataset for this masters thesis there are 126, as every participant with a missing value has been taken out. As students

start learning English at six years old, it is assumed that all participants had the same experience of 11 years total. While for German some may have started in lower secondary school, thus the amount of years learning German were 1, 2, 4 or 5 years. Participants were located in two cities from several classes when taking the tests. The last grade and self-rating of their own skill level were also provided in the dataset, but are not be used in the analysis or results in this master thesis.

The AJT for English and German consisted of 48 sentences, where 24 were the target sentences. These consisted of 12 sentences with topicalization, where the verb appearing in second position (V2) for 6 of them being grammatical, and verb appearing in third position (V3) for 6 of them being ungrammatical. A sentence with topicalization means that the sentences starts with a word that puts an emphasis on the topic or focus, with words like jeden/every. The other 12 sentences consisted of sentences with sentence adverbials, where the verb would precede or succeed the sentence adverbial for 6 of them respectively. A sentence adverbial is described as a word that give more information about a verb, like immer/always for L3/L2. The participants judge if a sentence is grammatical or ungrammatical on a 1-4 Likert Scale, where 1 is judging the sentence to be "Bad" and 4 being "Good", which is what the data consists of.

In this work we follow Dahl (2020) and the focus on sentences with topicalization and sentence adverbials. Similar to Dahl (2020) we investigate the research questions presented in Chapter 1.1.

With English sentences typically being a V3 language, the sentences with topicalization and sentence adverbials with V3 are considered to be grammatical. Sentences with V2 are thus considered to be ungrammatical. For the AJT it is expected that the participants will judge V3 English sentences with Likert score close to 4 and V2 English sentences close to 1. For German sentences, this is the complete opposite with V2 being grammatical and V3 being ungrammatical similar to L1 Norwegian.

Of the remaining 24 sentences, there are 12 that are considered filler sentences and 12

sentences presented as questions. These are also divided into 6 grammatical and 6 un-grammatical sentences, where the ungrammatical typically will miss the verb or lack an argument. With this in mind, it is likely easier to correctly classify the sentences as acceptable or not. The AJT in English and German were provided with filler sentences and questions following the same framework.

After we use the method described in Chapter 2 and find the proportion correct answers for the various sentence types, we present them using boxplots. Boxplots gives descrip-ton of how the data is distributed.
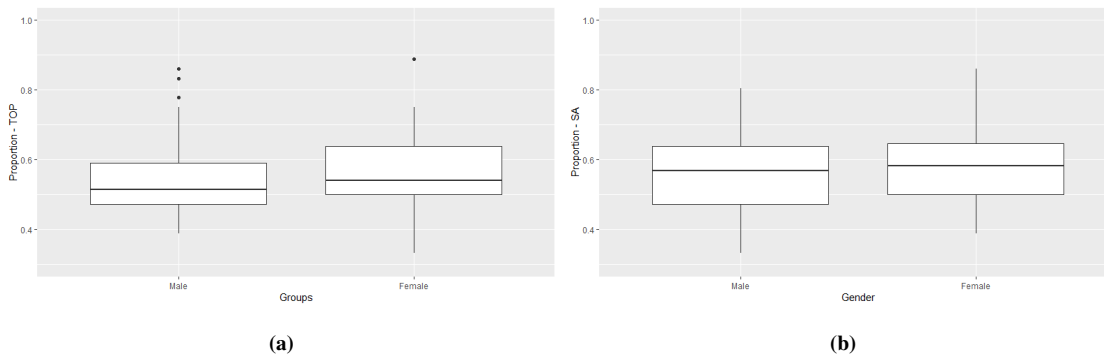
The boxplot in Figure 1 presents the data divided into the different YOG for the par-

(a)   (b)
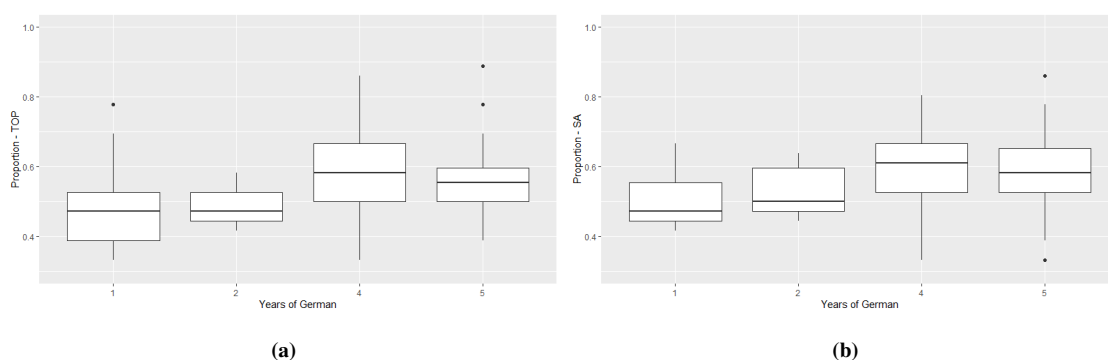
Figure 1: Boxplots for distribution of proportion correct judged German sentences with (a) topicalization and (b) sentence adverbial of different years of German.

ticipants. We observe that the median line inside the box is in general higher for the more experienced participants with 4 or 5 YOG. For both 2 YOG and 4 YOG the data is distributed within the upper 25% and the lower 25 % quartile. We observe that for 4 YOG the whiskers are long, which means that participants in this age group seems to have different opinions on a correct sentence, while 2 YOG got short whiskers and participants agree with each other more. We also note that for 2 YOG in both plots and for 1 YOG in Figure 1(a) the median is in the lower half of the box. This indicates that the data is right-skewed, or positively skewed.
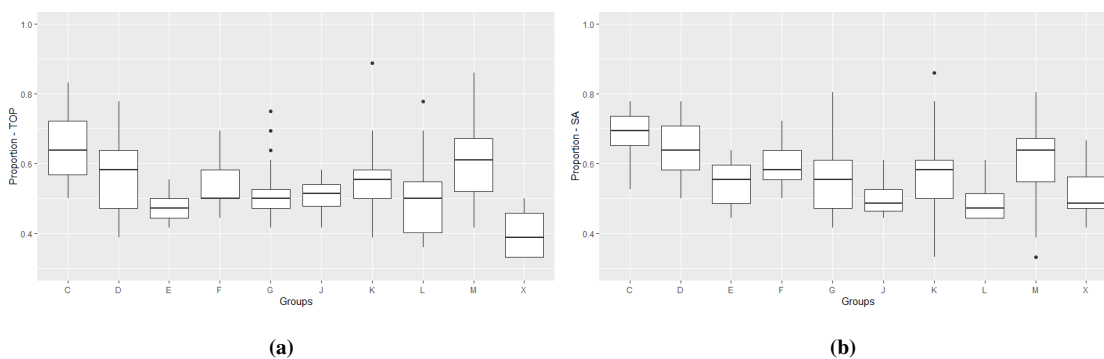
Boxplot in Figure 2 shows the data distributed for genders. There seems to small

(a)　　　　　　　　　　　　　　　　(b)

Figure 2: Boxplots for distribution of proportion correct judged German sentences with (a) topicalization and (b) sentence adverbial of different genders.

differences in distribution of proportion correct answer for German sentences with topicalization and sentence adverbial. The data tends to be positively skewed.

In Figure 3 we have a boxplot presenting the data distributed over the different groups

(a)　　　　　　　　　　　　　　　　(b)

Figure 3: Boxplots for distribution of percentage amount correct judged German sentences with (a) topicalization and (b) sentence adverbial of different groups.

that participants did the AJT in. We see that the data is distributed different for all the groups. Especially groups on the far left and far right in Figure 3(a) seems to be in disagreement of correct sentences.

In Figure 4 we have the distribution of proportion correct answer for German filler sentences and question sentences ungrouped. The median for WhGer seems to indicate that participants judge those sentences better than FillGer.
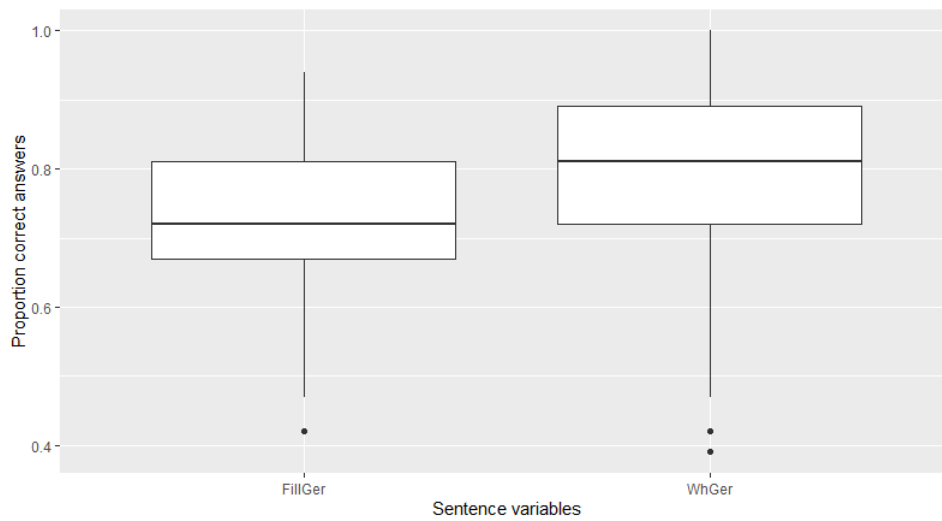
Figure 4: Boxplot for German filler sentences and question sentences, without grouping.

Figure 5 presents the distribution of proportion correct answer for English sentences.
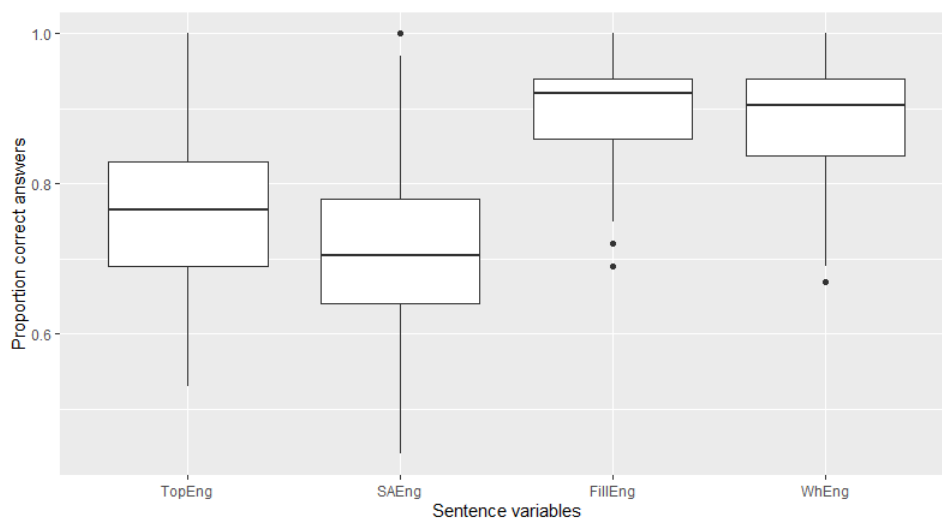
Figure 5: Boxplot for distribution of proportion correct answers for English sentences, divided into type of sentence.

They are grouped for the different sentence types. We observe that the participants judge English filler and question sentences much better, while also being more in agreement. We see this from the size of the inner quartiles and length of whiskers, which is

respectively smaller and shorter than for the English sentences with topicalization and sentence adverbial counterparts. Filler and question sentences are also left skewed.

3.2   Explanatory analysis

In this part we do an explanatory analysis of the data for each variable used and some analysis targeted to explore the interaction related to the research questions in Chapter 1.1.

In Figure 6 we observe the correlation between the different groups of sentences and their corresponding correlation coefficient. If two variables correlate with each other it means that a change in one of them also changes the other. A correlation coefficient of 1 means that they fully correlate. This is the case for the correlation between similar variables (TopGer - TopGer).

We observe that the correlation is positive for most of the variables, apart from a slight negative correlation between TopGer and TopEng. This means that a change in either of them has a negative effect. In terms of transfer this may indicate that an L2 transfer happens between these two type of sentences. The correlation is important for the regression models in Chapter 5. Preferably we want as little correlation between explanatory variables as possible to avoid overfitting the model. Between response and explanatory variables it is preferred that the correlation is higher.

We want to explore the data to see if there is any evidence of transfer between L2 and L3. We consider different representations of the data to see if we can find a trend in judgement that help us answer the research questions. In Figure 7 through 11 we have plotted the proportion of judgement scores chosen by all the participants, and grouped with years of German (YOG) for German and English sentences. We want to look for evidence of L2 transfer to L3 for the participants and judge whether proficiency in L2 structure influence transfer. We explore the same grammatical/ungrammatical sentences
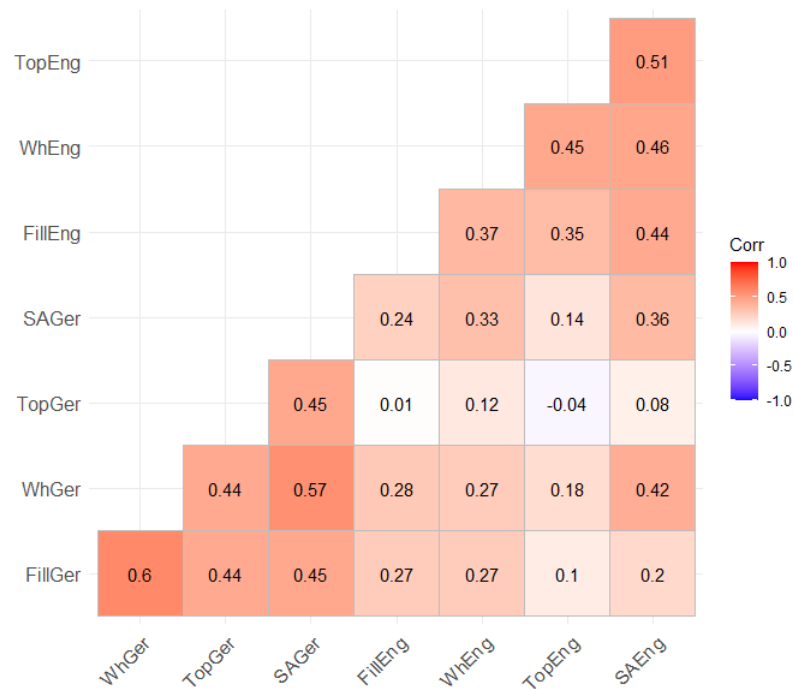
Figure 6: Correlation matrix with corresponding values of their coefficients

with topicalizations and sentence adverbials side by side for each language, while the filler and question sentences will be presented as a combined graph. The data will also be presented as a linear combination of the same type of sentences in German and English in Figure 12.

Figure 7 shows the grammatical sentences with topicalization. For these sentences we want a high proportion for Likert score of 3 or 4 chosen. Figure 7(a) shows a tendency towards 3 or 4 for topicalized German sentences. Participants with 4 or 5 YOG got the highest proportion of 3 or 4, which we assume is because of the increased experience with learning German. For participants with 1 year learning German there is tendency towards 3 and 4. There may be evidence of an L2 structure transfer, since we have a high proportion of judgement scores for 3 or 4 for topicalized English sentences. Participants with 4 or 5 YOG score high on English sentences, but on German sentences
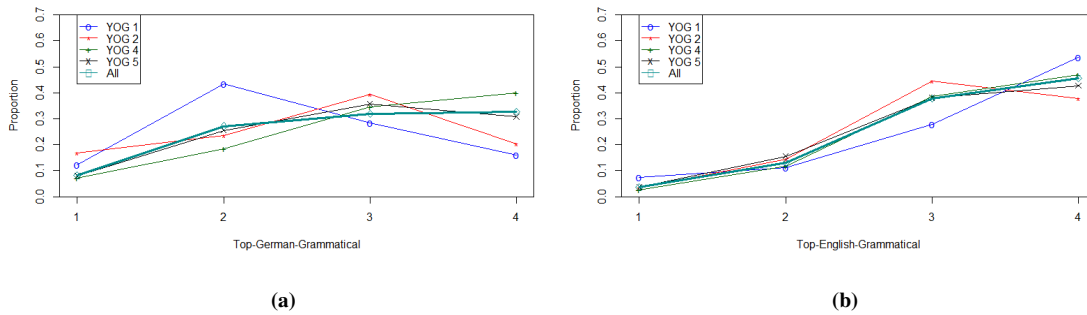
Figure 7: Percentage chosen judgements for grammatical topicalized sentences in (a) German and (b) English

they got less proportion 3 or 4 judged. We assume some L2 transfer may have resulted this. Participants with 2 YOG do score high on the English sentences, with the highest proportion of 4, but got the lowest porportion for German sentences. This suggests an L2 transfer as shown in Figure 7.
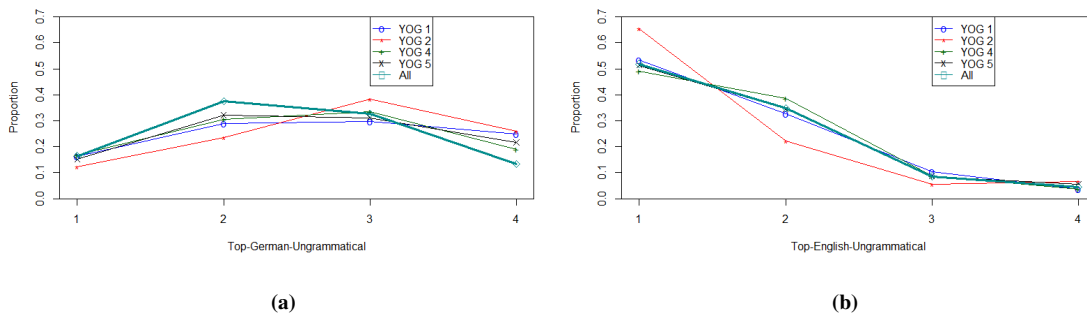
Figure 8: Proportion chosen judgements for ungrammatical topicalized sentences in (a) German and (b) English.

Figure 8 shows the proportion of judgement scores for ungrammatical topicalized sentences, with high proportions of 1 or 2 being the expected. We notice that a majority chose 2 or 3, which can indicate that the participants are not sure about the acceptability. Judgements of English sentences indicate that participants have a high proficiency

in English, which is the same for all years of learning German. There is not much difference for the judgements of German sentences. This suggest an L2 transfer, when RQ1 and 2 are considered.

The following plots show the distribution of Likert scores chosen for sentences with sentence adverbial.

(a)   (b)
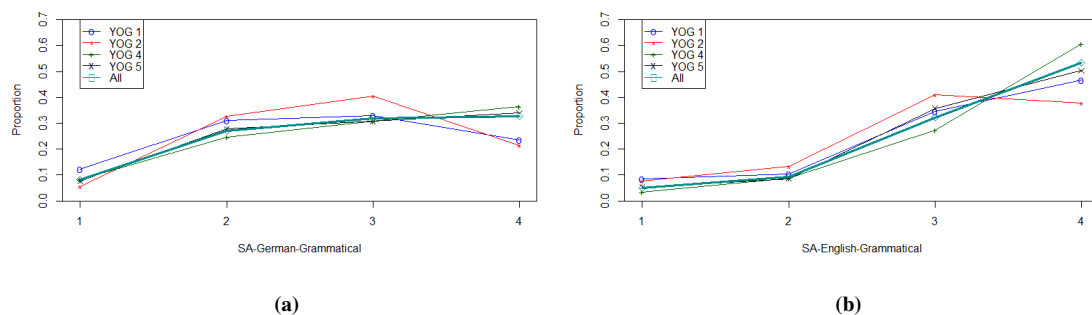
Figure 9: Proportion chosen judgements for grammatical sentences with sentence adverbial in (a) German and (b) English

Figure 9 shows the chosen judgements for grammatical sentences with sentence adverbial. The highest proportion chosen for German sentences are for scores of 2 or 3, which indicates that they are not sure about the acceptability for German sentences. Consider RQ1, and this development can be explained as being L2 transfer, when there seems to be less insecurity for the English sentences.

Figure 10 shows the ungrammatical sentences with sentence adverbial. We observe that for German sentences in Figure 10(a) the proportion chosen scores are much higher for score of 2, but also not a lot of difference between the proportion judging 1 or 4. For English sentences in Figure 10(b) there is a more definite difference between 1 or 4, which may indicate that there is an L2 transfer. Reason for this is the proportion chosen
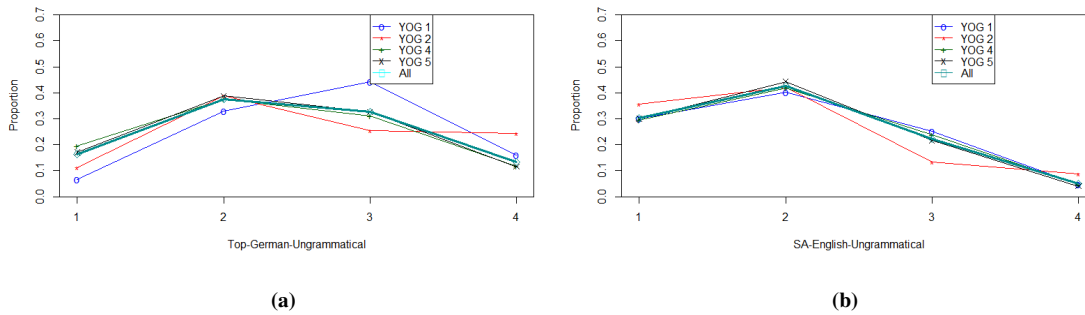
Figure 10: Percentage chosen judgements for ungrammatical sentences with sentence adverbial in (a) German and (b) English.

scores for ungrammatical German sentences, where we in this case have an L2 structure of V3 placement. The improved proportion for English sentences indicates that the better knowledge of the language is more important than the difficulty of the sentence structure.
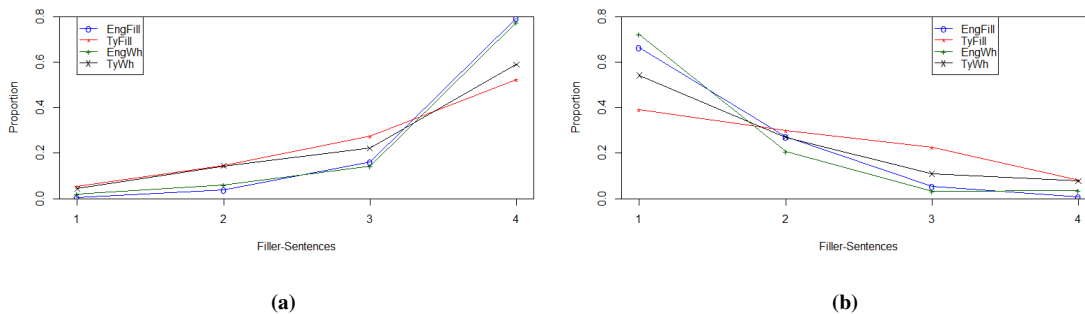
Figure 11: Percentage chosen judgements for filler sentences in L2 English and L3 German for (a) grammatical and (b) ungrammatical sentences.

The remaining variables are presented with all participants combined for proportion Likert score judged in Figure 11. These are the filler and question sentences in English and German. These sentences are considered obvious for the participants to judge whether they are grammatical or not, as mentioned in Chapter 3.1.1. As expected we observe
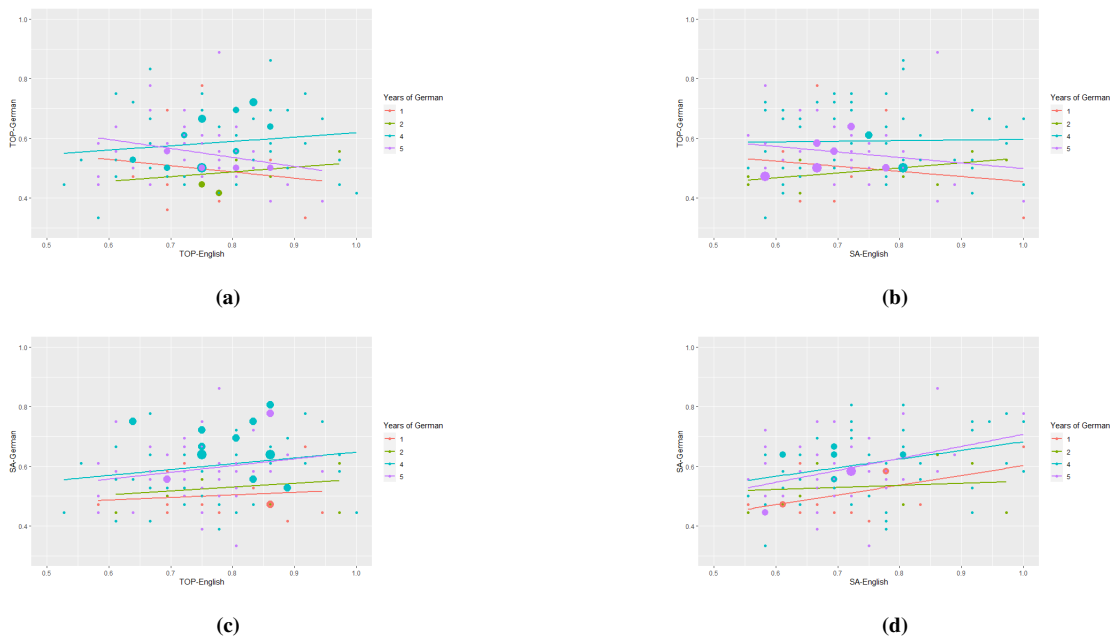
Figure 12: Linear relationship between sentences in German and English for (a) TopGer - TopEng, (b) TopGer - SAEng, (c) SAGer - TopEng and (d) SAGer - SAEng

that the graphs go towards 4 for the grammatical sentences and 1 for the ungrammatical sentences. We do notice that the participants seem to have an easier task judging English sentences than German sentences, by a slight margin. There is no evidence for L2 transfer in these plots, since they both follow the expected proportion chosen.

We want to determine if there is any L2 transfer when we introduce a linear relationship between the target sentences in German and English. The linear relationship is shown in Figure 12. Coefficients and p-values are in Table 13 in Appendix A.

The plots show a scatter plot with points of varying sizes. A big circle represents a point where there are many overlapping points. A simple linear model has been fitted for each group of YOG. We are estimating the proportion correct answers.

In Figure 12(a) we see the effects between topicalized sentences. We observe that for increasing proportion correct judged English sentences, we have increasing proportion

correct judged German sentences for participants with 2 or 4 years of learning German. There is a case to say that L2 transfer does not happen for 2 and 4 years of German, but this may not be because of proficiency and should be investigated further. As for 1 or 5 years of learning German, there is a negative relationship, and we can assume that there may be some transfer from L2. Reason for this being that the higher proportion correct answers of English topicalized sentences the participants with 1 or 5 YOG have, the lower proportion correct answers for German sentences with topicalization they have. The decrease indicates an L2 transfer, while there is less transfer for 2 or 4 YOG.

Figure 12(b) shows the linear relationship between German sentences with topicalization and English sentences with sentence adverbial, and we observe the same effects as for Figure 12(a). Again we assume there is an L2 transfer for 1 or 5 YOG.

Figure 12(c) shows the linear relationship between German sentences with sentence adverbial and English sentences with topicalization. We observe an increase in estimate proportion correct answers for SAGer when TopEng is increasing. This indicate less L2 transfer.

Figure 12(d) shows the linear relationship between German sentences with sentence adverbial and English sentences with sentence adverbial. We observe the same effect as for Figure 12(c), and assume less transfer.

4   Background: Statistical Methods

4.1   GLM

A Generalized Linear Model, or GLM, is a broad statistical framework that covers more than the assumed gaussian distributed models (Fahrmeir et al., 2013). We also want to model more than continuous variables, such as binary, categorical or count variables. In GLM we investigate these further by using different types of models while still working under the assumption that the effects of the covariates can be modelled through a linear predictor(Fahrmeir et al., 2013).

In our case we will model our response, the percentage of correctly judged sentences, as binomial distributed. The explanatory variables are also percentage correctly judged sentences.

In our dataset we got n individuals each given the form $(y_i, x_{i1}, \cdots, x_{ik})$, where k is the number of explanatory variables in the model and $i = 1, \cdots, n$. Here y_i takes the binomial responses in the interval $[0, 1]$, where 0 translates to "Failure" and 1 is "Success". In this paper we want to model the probabilities for "Success", and thus the estimate effects of the explanatory variables on the probability,

$$\pi_i = P(y_i = 1) \tag{2}$$

Here the response variable y_i is assumed individual for every participant. In this equation we can not ensure that π_i avoids exceeding the interval $[0, 1]$, which is impossible for probabilities. To avoid this we assume the model,

$$\pi_i = F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \tag{3}$$

and restricts the domain of F to $[0, 1]$.

Consider an additive model, then the linear predictor becomes (Fahrmeir et al., 2013; Dobson & Barnett, 2008)

$$\eta_i = \beta_0 + \beta_1 x_{i1} \cdots + \beta_k x_{ik} \tag{4}$$

with the covariates x_i and the regression coefficients $\beta_0, ..., \beta_k$. All x_{ik} here must take a value in the interval $[0, 1]$, when x_{ik} are categorical variables they can only take binary values 0/1. The restrictions β_i are described as challenging to handle (Fahrmeir et al., 2013). An example is that the predictor η_1 can surpass the interval $[0, 1]$. To manage this we need to express the predictor in Eq. 4 on the form,

$$\eta_i = g(\pi_i) \tag{5}$$

and to handle the restrictions on the parameters β_i use the relation combined with the probability function,

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \tag{6}$$

This ensures that $h(\eta_i)$ is in the interval of $[0, 1]$ as well, and takes away some of the restrictions for the parameters β_i. We combine the expression from Eq. 5 with Eq. 6 to show that,

$$g = h^{-1} \tag{7}$$

From this we can derive the link function that we want to estimate the linear predictor with. A logistic distribution as described in (Fahrmeir et al., 2013), is given as,

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \tag{8}$$

where the expression is monotonically increasing. By rearranging, we get the odds function,

$$\exp(\eta_i) = \frac{\pi_i}{1 - \pi i} \tag{9}$$

The odds function translates to the probability of "Success" over the probability of "Failure", like so

$$\frac{\pi_i}{1 - \pi i} = \frac{P(y_i = 1)}{P(y_i = 0)} \tag{10}$$

Which gives us the desired link function

$$\eta_i = \log\left(\frac{\pi_i}{1 - \pi i}\right) \tag{11}$$

This is called a logarithmic odds function, or a logit link function (Fahrmeir et al., 2013).

4.2   Model Fit

There are many methods for evaluating model fit, and in this section we present the ones used in this thesis. Most of the metrics used balance small discrepancy between fitted model and observations and overfitting. When a value is significant we say that the explanatory variable is influential on the model (Fahrmeir et al., 2013). An explanatory variable is significant if the p-value is less than 0.05. A non-significant variable is not necessarily without explanatory power, but that the estimate may have been random.

AIC, or Akaike Information Criterion, is a widely used criteria for model fit, but are generally only applied when comparing two different models. It uses the maximum value of the log-likelihood to calculate the value. AIC is defined by,

$$AIC = -2\ell(\hat{\beta}) + 2p$$

where ℓ is the log-likehood function with inputs estimated regression coefficients $\hat{\beta}$. p is the number of total residual coefficients in the model. When comparing AIC of two models we want the model with smallest value.

4.3   Methods for model choice

When choosing the best model for the response there are a lot of criteria in the selection process. All of the parameters mentioned in this section are viable when determining

model fit of the estimated response variable with explanatory variables. We want to explore the advantages of a stepwise selection. AIC of two models are compared, one with the full model and another where we eliminate non-significant covariates, and investigate if the new model is a better fit. The lowest AIC indicates the model with the best fit, and are compared in the next step when we eliminate another explanatory variable. We use the "stepAIC()" to do this procedure automatically using statistical software, and is well designed for this purpose (Zhang, 2016). What we end up with is a reduced model that has the lowest AIC, and gives the best fit with the explanatory variables for the response.

5   Results

In this chapter we fit the models to the data presented in Chapter 3.2 and interpret the results in terms of the research questions in Chapter 1.1. The results are based transformation from Likert scale to proportion correct answers, as pretested in Chapter 2. The model fit is analyzed using the methods presented in Chapter 4.2. An overview for the steps and AIC when using the stepwise selection method is presented in Appendix B.

5.1   Results simple GLM models for topicalization

First we consider the simplest version of the models. These models are on the form,

$$\eta_i = \beta_0 + \beta_1 x_{i1} \tag{12}$$

where x_{i1} is the explanatory variable of the model either being proportion correct judged sentences or a categorical variable. For our model we have "Years of German" (YOG) and "Gender" as categorical variables. We can use these to model proportion correct answers for sentences with topicalization or sentence adverbial considering effects of those groups only. As we have four groups of YOG, there are the model including only these variables is defined by

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

where x_{i1}, x_{i2} and x_{i3} are indicator variables for YOG=2, YOG=4 and YOG=5, respectively. The estimated proportion correct answer for 1 YOG is defined when all the covariates are 0, i.e. is accounted for in β_0. The model for estimating proportion correct judged sentences with topicalizations as response variable becomes,

$$\eta_i = -0.14 + 0.09 x_{i1} + 0.49 x_{i2} + 0.35 x_{i3} \tag{13}$$

Model in Eq. 13 predicts an increase in proportion correct answer for all participants. Notice that participants with 5 YOG have a lower coefficient value than 4 YOG, which

means that the model predicts better judgements for 4 YOG. This is comparable to what is presented in Figure 7, but not presented in Figure 8.

We model "Gender" with TopGer as response variable. This becomes

$$\eta_i = 0.17 + 0.06x_{i1}$$

Here the x_{i1} is 1 if gender is female. Predicted proportion correct judged sentences is better for female than male, but not significant. The difference here is very small, which is also observed in Figure 2. For this reason we eliminate this variable from the rest of the models.

In Table 1 the simple models for German sentences with topicalization (TopGer) as response variable with one sentence as explanatory variable are presented. C.I. is the confidence interval.

We observed that the effect of German sentences with sentence adverbial (SAGer)

Model	x_{i1}	β_0	β_1	C.I. (β_1)	p-value (β_1)
MT1	SAGer	-0.92	1.96	$[1.40, 2.52]$	~ 0
MT2	TopEng	0.37	-0.21	$[-0.81, 0.41]$	0.48
MT3	SAEng	-0.02	0.31	$[-0.21, 0.81]$	0.20
MT4	FillEng	0.16	0.05	$[-0.85, 0.95]$	0.91
MT5	WhEng	-0.40	0.69	$[-0.16, 1.45]$	0.07

Table 1: Simple models with topicalized German sentences as response. Green means significant (p-value < 0.05), red means not significant, cyan means close to significant (0.05,1).

and English question sentences (WhEng) are significant in their corresponding models, MT1 and MT5 respectively, while models MT2 through MT4 are not significant. In model MT1 the explanatory variable is the proportion correct answers for SAGer that has a positive effect on response variable TopGer. This means that proportion correct

judged TopGer increases the better a participant is at judging SAGer sentences. For MT5 we have the explanatory variable WhEng, which also has a positive effect on TopGer. Model MT3 and MT4 indicates an increase in proportion correct answers for TopGer with increasing proportion correct answer for the explanatory variables, but they are not significant. Reason for this can be the lack of correlation, as seen in Figure 6. In model MT2 we observe a non-significant negative relationship between explanatory variable TopEng and response TopGer. Model MT5 is close to significance. An overview of these simple models give us an impression on what explanatory variables that correlate and not.

5.2   Results simple GLM models for sentence adverbial

We look at German sentences with sentence adverbials (SAGer). As with sentences with topicalization, we first look at the effects that years learning German got on the proportion correct judgements. This model becomes,

$$\eta_i = 0.01 + 0.11x_{i1} + 0.41x_{i2} + 0.36x_{i3} \tag{14}$$

If we compare with Eq. 13 we observe that both are increasing for all categorical variables, and coefficient for 4 YOG is higher than for 5 YOG. This means that there is little difference between these two models, and both predict higher proportion correct judged response variable for 4 YOG than 5 YOG.

Model for "Gender" as categorical variable is given with,

$$\eta_i = 0.27 + 0.09x_{i1}$$

Model shows that proportion correct judged SAGer sentences for female is higher than for male, but not enough to determine that there is a difference between the genders as it is not significant. As for the same model with TopGer as response variable, we

eliminate this categorical variable from other models for the same reason.

Table 2 shows the simple models for one explanatory variable each, with SAGer as response variable and English sentences and TOPGer as explanatory variable.

We observe that all the models have significant values for the coefficient β_i. For MS1

Model	x_{i1}	β_0	β_1	C.I. (β_1)	p-value (β_1)
MS1	TOPGer	-0.66	1.79	$[1.25, 2.33]$	~ 0
MS2	TOPEng	-0.14	0.61	$[0.01, 1.22]$	0.04
MS3	SAEng	-0.64	1.34	$[0.84, 1.84]$	~ 0
MS4	FillEng	-1.16	1.66	$[0.56, 2.36]$	~ 0
MS5	WhEng	-1.34	1.88	$[1.12, 2.64]$	~ 0

Table 2: Simple models with sentence adverbials in German sentences as response. Green means significant ($p < 0.05$), red means not significant.

we notice a positive relationship between explanatory variable TopGer and response SAGer. We also have significance for the same model with switched positions in Table 1. This connection is a result of them correlating, as seen in Figure 6. For the models MS2 through MS5 that have explanatory variables with English sentences are all significant. Again, we can assume the reason is because of the positive correlation between SAGer and the explanatory variables, seen in Figure 6. Considering MS2, we see that the significance leads us to believe it is the worst fit.

5.3   Results model choice GLM for topicalization

In this section we are focusing on additive models. We increase the number of explanatory variables used and investigate models to estimate proportion correct answers for German sentences with topicalization (TopGer). We first focus on English sentences

only as explanatory variables, and find the best possible fit using the stepAIC function described in Chapter 4.3. This is further increased to look for the best fitting model using all explanatory variables in our data set.

The full additive model including all English sentences and the categorical factor Years of German (YOG) becomes,

$$\eta_i = -0.25 + 0.25x_{SAEng} - 0.39x_{FillEng} + 0.72x_{WhEng}$$
$$- 0.42x_{TopEng} + 0.07x_{2YOG} + 0.45x_{4YOG} + 0.32x_{5YOG} \tag{15}$$

where x_{4YOG} and x_{5YOG} are the only significant explanatory variables for proportion correct TopGer. We want to reduce this model by eliminating non-significant explanatory variables in a stepwise model choice method. This results in the model,

$$\eta_i = -0.14 + 0.09x_{2YOG} + 0.49x_{4YOG} + 0.35x_{5YOG}$$

which we already investigated in Eq. 13. This model is not benefitial for the analysis in this subsection, as we want to see the effect of other English sentences on the response variable TopGer. We eliminate all YOG variables from our full model in Eq. 15 and all upcoming models. We use the stepAIC function on the full model without YOG to find the best fitting model.

TopGer	Coefficient	C.I.	p-value
β_0	-0.26	$[-0.96, 0.44]$	
TopEng	-0.56	$[-1.22, 0.10]$	0.09
WhEng	1.01	$[0.17, 1.85]$	0.02

**Table 3:** Model for response variable German sentences with topicalization. Green means significant (p<0.05), cyan means close to significant (p $\in$ (0.05, 0.10)).

Our best model consists of the significant explanatory variables English sentences with topicalization (TopEng) and English question sentences (WhEng). The negative relationship between TopGer and TopEng may be a result of the negative correlation from

Figure 6. We also observed this in Table 1. Correlation is not big and the significance may be affected by this. Overall TopGer is not correlating much with the explanatory variables, being -0.04 for TopEng and 0.12 for WhEng, while the explanatory variables got a correlation coefficient of 0.45. We explore what this means in the next paragraph. The linear predictor function becomes,

$$\eta_i = -0.26 - 0.56x_{TopEng} + 1.01x_{WhEng} \tag{16}$$

where x_{TopEng} is the explanatory variable TopEng and x_{WhEng} is the variable WhEng. To investigate effects of the explanatory variables on the response we show the model in a plot.
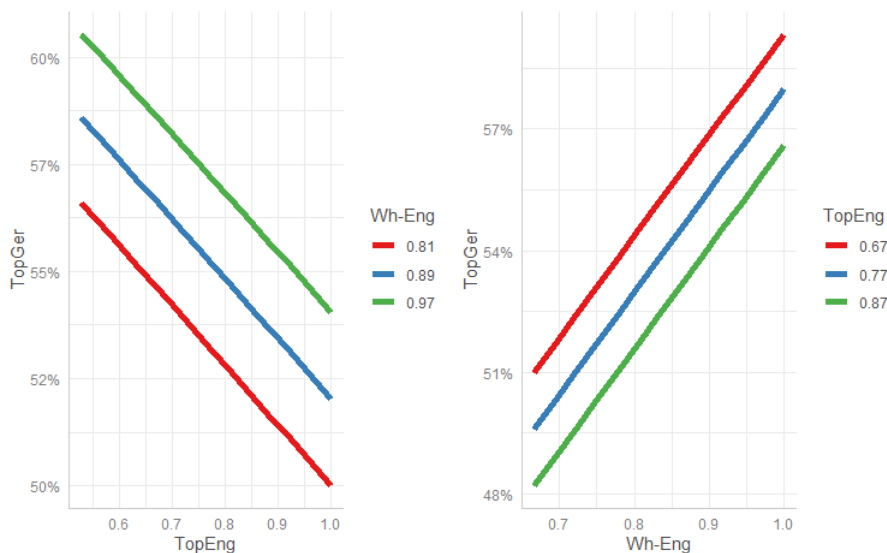
Figure 13: Plot for Eq. 16. Blue lines correspond to the estimated average of respectively WhEng and TopEng. Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

Figure 13 presents the model in Eq. 16 for both the explanatory variables. The left plot in Figure 13 we see that the relationship is negative when we gradually increase the proportion correct answers of TopEng. This means that for such an increase, there is a decrease in proportion correct answers TopGer. In the right plot we see a positive

relationship for WhEng, but there is still a negative relationship for TopEng. Reason for that is when we increase TopEng, the proportion of TopGer decreases, as seen by the lines corresponding to their respective value of TopEng. As seen in Table 1, there is a negative relationship when TopEng is the only explanatory variable for estimating proportion correct judgements of TopGer.

We increase available explanatory variables to include proportion correct answer for German sentences as well. We are using the stepwise model selection method to get the model with best fit for the response variable TopGer with the explanatory variables. When the full model is reduced we get the model,

TopGer	Coefficient	C.I.	p-value
β_0	-0.60	$[-1.41, 0.22]$	
SAGer	1.24	$[0.56, 1.91]$	~ 0
FillGer	1.04	$[0.38, 1.70]$	~ 0
FillEng	-1.26	$[-2.20, -0.33]$	0.01
WhGer	0.60	$[-0.01, 1.20]$	0.05

Table 4: Model for response variable German sentences with topicalization. Green means significant (p<0.05), cyan means close to significant (p $\in$ (0.05, 0.10)).

Notice that all explanatory variables for German sentences effects the response positively, while English filler sentences (FillEng) has a negative effect on the estimate. This effect may be explained that participants that got a high proportion of German sentences correctly judged also are assumed good at judging other German sentences. The negative effect from FillEng may be explained that a high proportion correct answers for English sentences leads to some transfer. We also see that the effect of German question sentences (WhEng) on the response is less significant than the other explanatory variables.

The corresponding linear predictor becomes,

$$\eta_i = -0.60 + 1.24 x_{SAGer} + 1.04 x_{FillGer} - 1.26 x_{FillEng} + 0.60 x_{WhGer} \qquad (17)$$

Which gives the plot,

We see that when the proportion correct answers of SAGer is increasing, the proportion
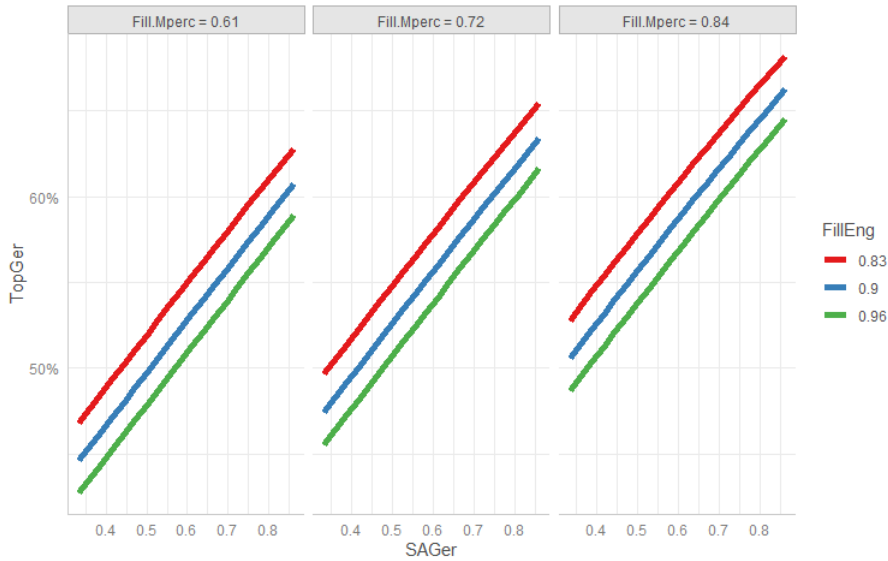
Figure 14: Plot for Eq. 17. Blue lines are the estimated average for FillEng, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

correct answer for TopGer is increasing as well. Notice that for increasing proportion correct answer of FillEng, proportion correct TopGer decreases. This effect may assume that even participants that are judging German sentences good still have some transfer if they are judging English sentences with high proportion correct as well.

5.4   Results model choice GLM for sentence adverbial

We change to German sentences with sentence adverbial (SAGer) as response variable. As with TopGer in the previous section, we use the stepwise model selection to decide

the best fitting model for the response. In the first part we focus on English sentences as explanatory variables only, and in the second part include all explanatory variables. We use the stepwise model selection method on the full model with SAGer as the response variable to get the model with the best fit, being,

SAGer	Coefficient	C.I.	p-value
β_0	-1.30	$[-1.99, -0.61]$	
SAEng	1.17	$[0.57, 1.77]$	~ 0
WhEng	1.39	$[0.52, 2.27]$	~ 0
TopEng	-0.58	$[-1.30, 0.13]$	0.11

Table 5: Model for German sentences with sentence adverbial as response variable. Green means significant (p < 0.05), red means not significant.

Table 5 shows the model for proportion correct SAGer as response variable, English sentences with sentence adverbial (SAEng), topicalization (TopEng) and English question sentences as explanatory variables. We observe that TopEng is not significant for the model. If we look at the correlation matrix in Figure 6 we see that all the explanatory variables got some correlation with each other. For SAEng and TopEng we have a coefficient of 0.51, which means that a change in one of them will change the other. The response variable SAGer do correlate with SAEng and WhEng with 0.36 and 0.33 respectively, which is preferred. Together with the significant p-value we assume that they are good explanatory variables for this model. TopEng got a coefficient of 0.14 with SAGer. A low correlation with the response means that a change in the TopEng has lower effect on SAGer than a change in WhEng or SAEng have. This may be a reason to eliminate TopEng from the model, indicating that the higher correlation with the other explanatory variables may be overfitting the model. We choose to keep it for exploratory reasons and comparison. The significant variables WhEng and SAEng are

both positive, which indicate that an increase for their proportion correct answers will increase proportion correct SAGer. The negative effect of TopEng on SAGer is not significant, and we consider it not effecting the estimate much. The linear predictor for the model becomes

$$\eta_i = -1.30 + 1.17x_{SAEng} + 1.39x_{WhEng} - 0.58x_{TopEng} \tag{18}$$

with the corresponding model plot,
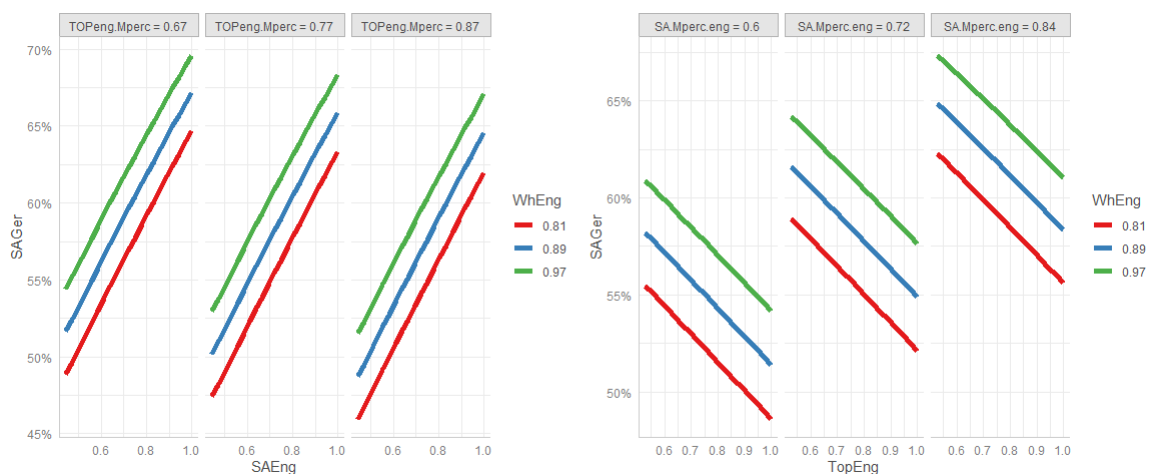
Figure 15: Plot for Eq. 18. Blue lines are the estimated average of WhEng, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

Plots in Figure 15 shows the model in Eq. 18. We observe that the proportion correct answers for SAGer is increasing for every increase in SAEng and WhEng, while increase in proportion correct answer for TopEng has an effect of a slight decrease in proportion correct answers SAGer. There seems to be a negative effect from judging TopEng sentences correct on the German sentences in focus, as is also observed in Chapter 5.3. The corresponding SAEng variable has a positive effect, and a significant effect on SAGer as response. This is expected because of the higher correlation, as stated above.

We increase explanatory variables to include German sentences. The stepwise model

selection is used to decide on the best additive model to estimate proportion correct answers for German sentences with sentence adverbial (SAGer). The stepwise model selection give us the best fit from the full model,

SAGer	Coefficient	C.I.	p-value
β_0	-2.26	$[-2.97, -1.55]$	
TopGer	1.06	$[0.47, 1.64]$	~ 0
WhGer	1.22	$[0.68, 1.75]$	~ 0
WhEng	0.83	$[-0.02, 1.75]$	0.06
SAEng	0.44	$[-0.02, 1.69]$	0.15

Table 6: Model for German sentences with sentence adverbial as response variable. Green means significant (p < 0.05), red means not significant.

Table 6 shows the coefficients and significance of the explanatory variables for the best fitting model. We observe that all explanatory variables have a positive effect on the response. This means that if the proportion correct answer for any of the variables increase the predicted estimate of proportion correct answers for SAGer increase as well. The model consists of both German and English sentences as explanatory variables, meaning that proficiency in both English and German have an effect on proportion correct SAGer sentences. We get the predictor,

$$\eta_i = -2.26 + 0.44x_{SAEng} + 0.83x_{WhEng} + 1.06x_{TopEng} + 1.22x_{WhGer} \qquad (19)$$

With the corresponding plot,

Figure 16 represent the model in Eq. 19. We notice that the response SAGer is increasing when every explanatory variable is increasing, and vice versa for decrease. This is in agreement with what we observe in the model Table 6.
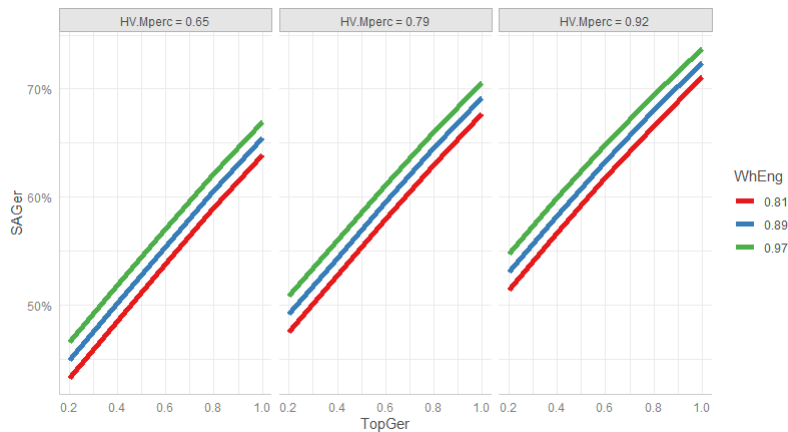
Figure 16: Plot for Eq. 19. Blue lines are estimate average of WhEng, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

5.5 Results models with interactions

We have looked at additive models for response variables TopGer and SAGer. With response variable TopGer we had a best fit model with English sentences only as explanatory variables in Table 3 with a negative effect from TopEng that had a p-value of 0.09, and had less significance for the model than WhEng in the same model. In Table 4 the best fit model with all sentence variables showed that WhGer had less singificance in the model. For SAGer we found a best fit model with only English explanatory variables that had TopEng not signficant for the model in Table 5, and a best fit model for all explanatory models in Table 6 with SAEng not significant and WhEng less significant for the model. We want to explore if there is an interaction between the variables in the best fit model with a significant effect on the model. Such an interaction can improve the model to predict better estimates of the proportion correct answers for the response variable.

Models in this section are decided in possible two methods. We can choose the already best fit for the additive models with an interaction between two or more of the existing explanatory variables. We get a model that looks like

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

The other method is to use the stepwise model choice method using the statistical software presented in Chapter 4.3. We end up with two models, one that may show an improvement on the already best fit additive model and another that may show a different improved model.

5.5.1   Interaction models with response German topicalization

From the model in Eq. 16, we want to see if we can improve on the model with an interaction term. We let TopEng and WhEng in interact, and get the model,

TopGer	Coefficient	C.I.	p-value
β_0	-7.84	$[-12.87, -2.82]$	
TopEng	9.78	$[2.96, 16.61]$	~ 0
WhEng	9.52	$[3.88, 15.18]$	~ 0
TopEng $\times$ WhEng	-11.55	$[-19.15, -3.97]$	~ 0

Table 7: Model for German sentences with topicalization as response variable with interaction. Green means significant ($p < 0.05$)

In Table 7 we observe a significant interaction coefficient for the English sentence explanatory variables TopEng and WhEng. We see that the interaction has a negative effect on the response TopGer, which means that a participant with high proficiency in English may do worse on judging German sentences with topicalization. We observe in Figure 6 that TopEng and WhEng got a correlation coefficient of 0.45, which means much of the data in those variables correlate. This may indicate that the model in Table 7 is overfitted and may not be an improvement to the additive model. The high values

for the estimated coefficients in the model, and the size of the confidence interval of the estimates are also indications that the model is overfitted. When a model is overfitted the model picks up all the noise in the model and tries to perfectly fit all the observed data rather than predict them.

The predictor for the model becomes,

$$\eta_i = -7.84 + 9.78x_{TopEng} + 9.52x_{WhEng} - 11.55x_{TopEng}x_{WhEng} \tag{20}$$

and the plot for the model becomes

We observe the effects of the interaction term when participants judge English sen-
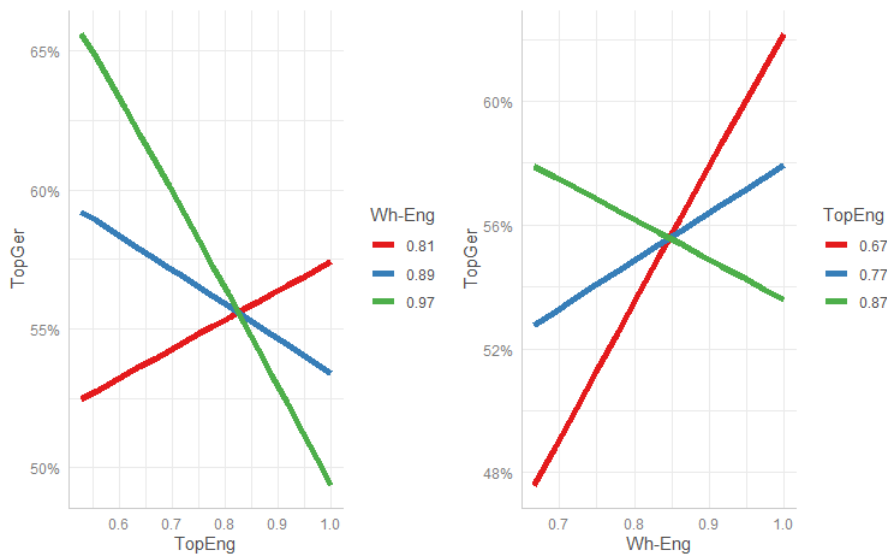
Figure 17: Model for TopGer as response variable with interaction TopEng×WhEng. Blue lines are the estimated average for WhEng and TopEng respectively, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

tences with a high proportion correct, showing increased proficiency. The green line, representing a high proportion correctly judged answers for WhEng, predicts that for increased proportion correct judged TopEng, we get a decrease in proportion correct TopGer. However, with a decrease in proportion correct for either TopEng or WhEng

seems to have a positive effect on proportion correct TopGer, and a participant that judges either TopEng or WhEng with low proportion correct get a higher proportion correct answers for TopGer with increasing proportion correct of the other variable. This is observed in Figure 17 when WhEng or TopEng is low, and the other high, we observe a higher proportion correct answers for TopGer.

We use stepwise model choice looking for the best fit from a full model. This becomes,

TopGer	Coefficient	C.I.	p-value
β_0	-12.15	$[-17.81, -6.51]$	
TopEng	6.27	$[-1.46, 14.02]$	0.11
SAEng	10.46	$[3.36, 17.57]$	~ 0
WhEng	13.97	$[7.72, 20.25]$	~ 0
TopEng $\times$ WhEng	-7.78	$[-16.51, 0.94]$	0.08
SAEng $\times$ WhEng	-10.97	$[-18.84, -3.11]$	0.01

Table 8: Model for German sentences with topicalization as response variable with interactions. Green means significant (p < 0.05), red is not significant and cyan is close to significant (p $\in$ (0.05,0.10)

The stepwise model choice for best fit give us the model in Table 8. We observe that the best fit is increased with English sentences with sentence adverbial (SAEng) as explanatory variable and an extra interaction variable between SAEng and WhEng. Both of these are significant for the model, while the interaction term TopEng $\times$ WhEng that we had significant in Eq. 20 now is less significant for the model. The same goes for TopEng that is not significant for the best fit model in Table 8. The interactions have a negative effect on the response variable TopGer, but is only significant for the interaction between SAEng and WhEng. In Figure 6 we see that all of the explanatory variables do correlate some with each other. TopEng and WhEng with a correlation coefficient of 0.45 while SAEng and WhEng correlate with 0.46. This can be a reason to assume

some overfitting in this model.

The linear predictor becomes,

$$\eta_i = -12.15 + 6.27 x_{TopEng} + 10.46 x_{SAEng} + 13.97 x_{WhEng}$$
$$- 7.78 x_{TopEng} x_{WhEng} - 10.97 x_{SAEng} x_{WhEng}$$

(21)

And the corresponding plot for the function,
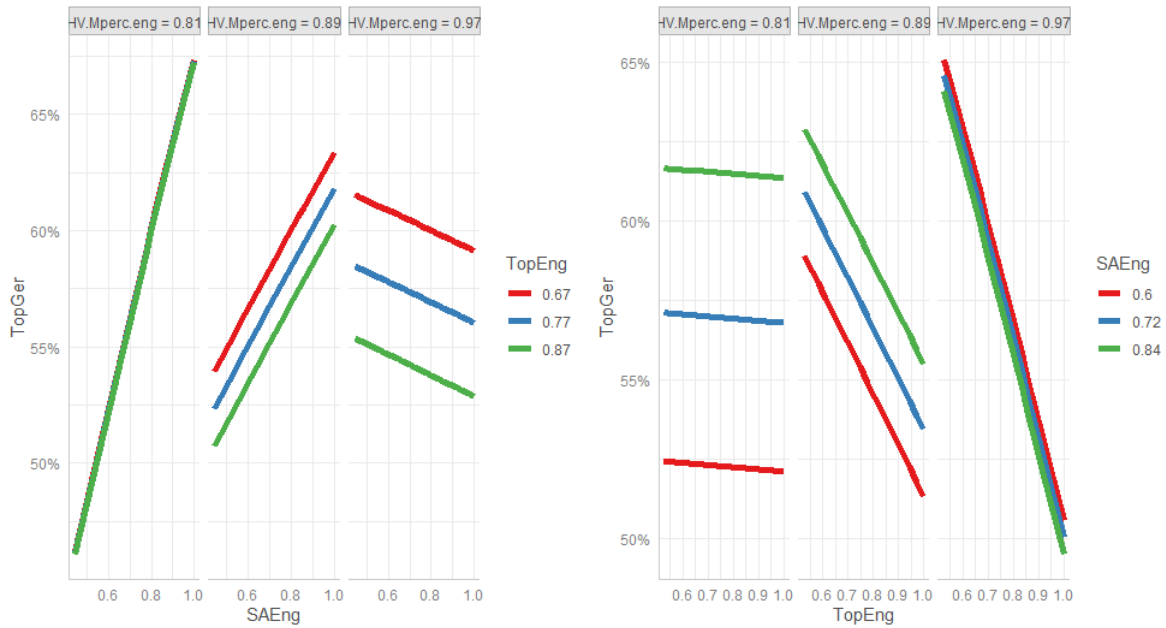
Figure 18: Model for TopGer as response variable with interactions TopEng×WhEng and SAEng×WhEng. Blue lines are the estimated average for TopEng and SAEng respectively, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

The plots in Figure 18 show a negative effect when we increase the proportion correct answers in the explanatory variables. The distance between the lines increase visually when we increase the value of WhEng. For an increase in proportion of WhEng the proportion correct judged TopGer decrease when SAEng increase. This is observed in the far right graph for the left plots in Figure 18. An increase in TopEng predicts a decrease in proportion correct TopGer. Combined we observe that when a participant got

high proportion correct judgements of English sentences, they have a lower proportion correct judgements of TopGer.

The right plot in Figure 18 shows a decrease in proportion correct TopGer when proportion of TopEng is increased. An increase in proportion correct WhEng has the same effect, while an increase in SAEng seems to predict higher proportion correct answers for the response TopGer by increasing the distance between the lines. There is a significant negative effect when both TopEng and WhEng are increased.

Models with English sentences as explanatory variables only seem to predict a decrease in proportion correct answer for German sentences with topicalization when participants show proficiency with English sentences. This is in this section shown using interaction, where we see the interaction change fast for varying proportion correct answers of English sentences in Figure 17, while in Figure 18 we see more parallell lines with the distance between them changed as the proportions are decreased or increased.

We investigate models where all explanatory variables are used to determine if there is an effect from interaction term. Using the stepwise model choice method do not give satisfactory results for a best fitting model. When all sentence variables are included to find a model, the stepAIC function return the full model without elminiating any explanatory variables and the interactions between them. To circumvent this we run stepAIC using the explanatory variables from Table 4 with interaction to find the best fitting model.

Table 9 shows the best fitted model. We see a negative effect from the interaction between explanatory variables English filler sentences (FillEng) and German question sentences (WhGer) and for the explanatory variable German filler sentences (FillGer). From Figure 6 we see that FillGer have a correlation coefficient of 0.45 with the response TopGer. As we aim to include explanatory variables that can explain the response variable, this is preferable. FillEng and WhGer have a low correlation coeffi-

TopGer	Coefficient	C.I.	p-value
β_0	-3.52	$[-7.39, 0.34]$	
SAGer	1.23	$[0.55, 1.91]$	~ 0
FillGer	-4.60	$[-7.48, -1.73]$	~ 0
FillEng	6.32	$[1.47, 11.20]$	0.01
WhGer	4.14	$[-0.96, 9.27]$	0.11
FillGer $\times$ WhGer	7.33	$[3.63, 11.04]$	~ 0
FillEng $\times$ WhGer	-9.63	$[-15.89, -3.43]$	~ 0

Table 9: Model for German sentences with topicalization as response variable with interactions. Green means significant (p < 0.05), red is not significant

cient of 0.28, which we consider good between explanatory variables. Because of these reasons, we assume that these explanatory variables have the most explanatory power for this model.

$$\eta_i = -3.52 + 1.23x_{SAGer} - 4.60x_{FillGer} + 6.32x_{FillEng} + 4.14x_{WhGer} \\ + 7.33x_{FillGer}x_{WhGer} - 9.63x_{FillEng}x_{WhGer} \tag{22}$$

The plot for model in Eq. 22 becomes, Figure 19 present the model in Eq. 22. In the left plot we have the predicted values for proportion correct judged TopGer when SAGer is increasing and FillEng is varied. We notice that for higher proportion correct answers of FillEng, the proportion correct answer of TopGer decreases. Since the explanatory variable FillEng has a positive effect on the model, then we assume this is the negative effect from the interaction FillEng $\times$ WhGer. We see that the positive effect from the interaction FillGer $\times$ WhGer is present in the right plot, where for higher proportion correct FillGer we have an increase in proportion correct TopGer. Since FillGer have a negative effect on the model alone, we assume that the interaction has a stronger effect on the model from this observation. We see that the proportion correct answers for
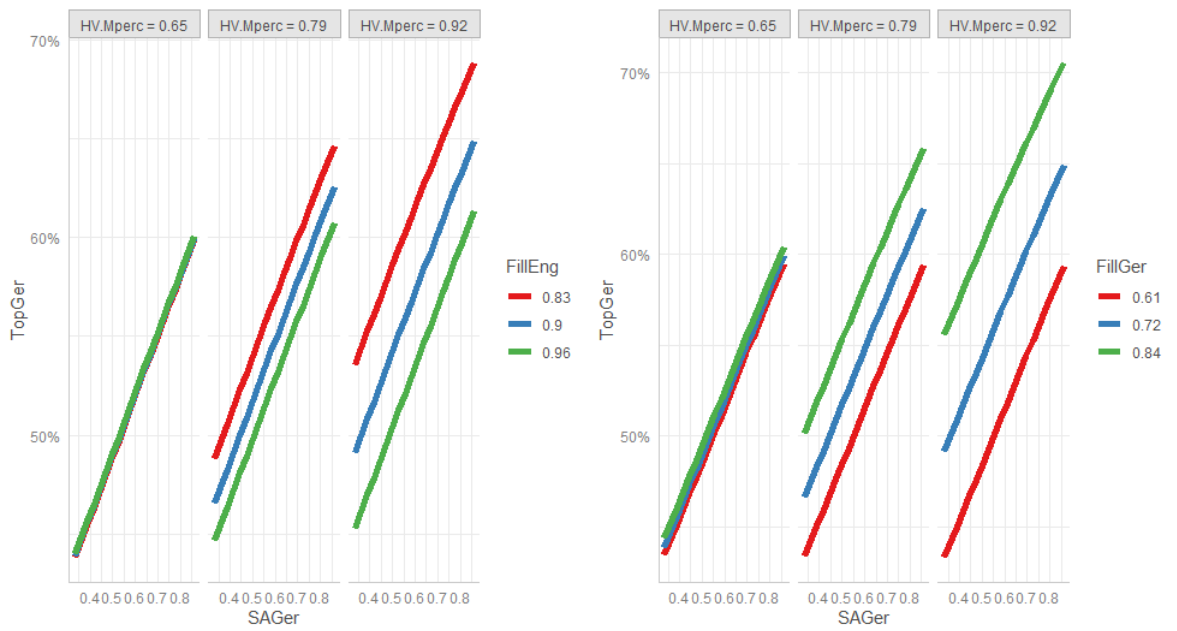
Figure 19:   Model for TopGer as response variable with interactions FillGer×WhGer and FillEng×WhGer. Blue lines are the estimated average for FillEng and FillGer respectively, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

TopGer is increasing when WhGer and SAGer are increasing in proportion.

In this section we observed that introducing an interaction term or more will have an effect on the model overall. The model in Figure 18 show a strictly negative effect from increasing proportion correct judged English sentences, while in Figure 19 we observe that proportion correctly judged English filler sentences have a negative effect on the response variable TopGer alone from the interaction.

5.5.2   Interaction models with response German sentence adverbial

We use the stepwise model choice function to find the best fitting model with English sentences as explanatory variables and interaction between them. Our model becomes,

SAGer	Coefficient	C.I.	p-value
β_0	3.53	$[-6.17, 13.26]$	
SAEng	3.74	$[0.59, 6.90]$	0.02
WhEng	-6.59	$[-17.45, 4.25]$	0.23
FillEng	-7.52	$[-18.62, 3.55]$	0.18
TopEng	1.83	$[-1.10, 4.76]$	0.22
SAEng $\times$ TopEng	-3.46	$[-7.37, 0.45]$	0.08
FillEng $\times$ WhEng	9.05	$[-3.36, 21.50]$	0.15

Table 10: Model for German sentences with sentence adverbial as response variable with interactions. Green means significant (p < 0.05), red is not significant, cyan is close to significant (p $\in$ (0.05,0.10)

The model in Table 10 is a reduced version using the stepwise model choice function stepAIC. We observe that the only significant explanatory variable is the English sentences with sentence adverbial (SAEng). It has a positive effect on the response variable, which means that an increase in proportion correct answers for SAEng predicts an increase in proportion correct answers for SAGer. While not significant, we notice that

WhEng and FillEng have a negative effect on SAGer. The interaction between SAEng and TopEng is close to significant, and has a negative effect on the response as well. In the correlation matrix in Figure 6 we have a high correlation coefficient of 0.51 between SAEng and TopEng. This can indicate that we are overfitting the model with explanatory variables that correlate with each other.

The model got the linear predictor

$$
\begin{aligned}
\eta_i = & 3.53 + 3.74 x_{SAEng} - 6.59 x_{WhEng} - 7.52 x_{FillEng} + 1.83 x_{TopEng} \\
& - 3.46 x_{SAEng} x_{TopEng} + 9.05 x_{FillEng} x_{WhEng}
\end{aligned}
\tag{23}
$$

The corresponding visual plot for model in Eq. 23 becomes
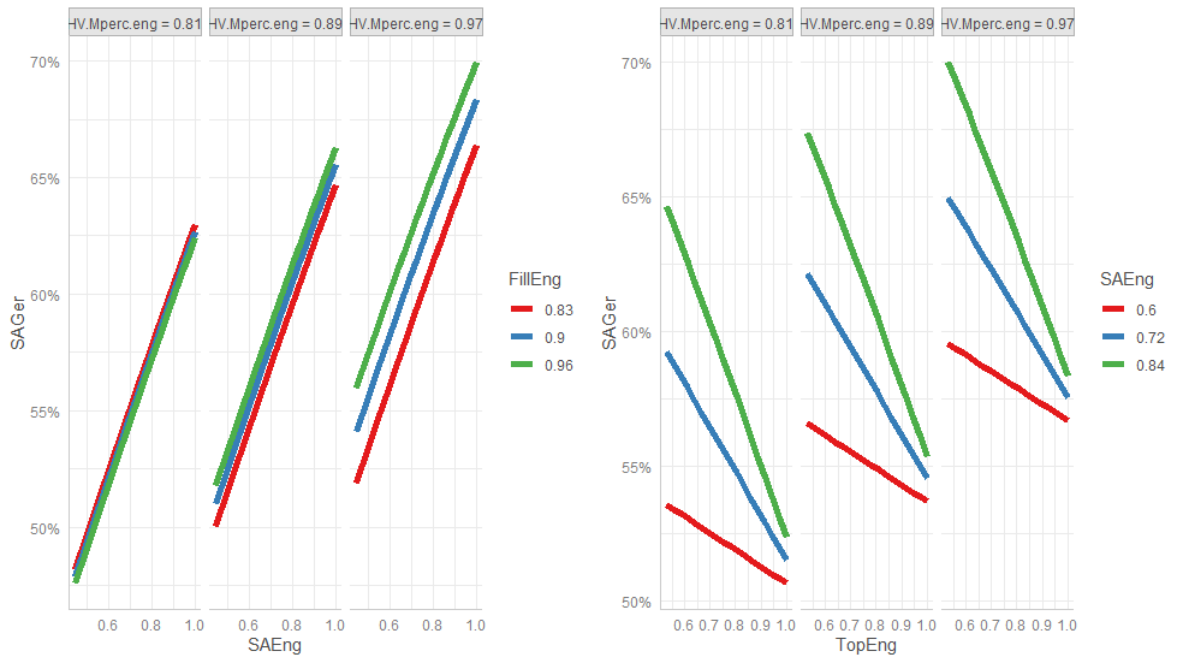
Figure 20:   Model for SAGer as response variable with interactions SAEng×TopEng and FillEng×WhEng. Blue lines are the estimated average for FillEng and SAEng respectively, Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

Figure 20 presents the model from Eq. 23. We observe in the right plot the decrease

in proportion correct answers for SAGer when the explanatory variables English sentences are increasing. We notice that TopEng and SAEng both have a positive effect on the model separately, so we assume that the negative effect on the model in Figure 20 is from the interaction between them. The increase in English question sentences (WhEng) have a positive effect on the response, and we observe an increase in proportion correct answers for SAGer in both plots. For increasing proportion correct FillEng we also observe an increase for SAGer. As both of them separately have a negative effect on the model we assume again that the effect from the interaction is bigger.

We include the German sentences as explanatory variables to check if there is an effect of adding an interaction to the model. We use the stepwise model selection adding all variables to estimate the response. The model becomes,

SAGer	Coefficient	C.I.	p-value
β_0	6.16	$[-4.46, 16.80]$	
SAEng	-3.24	$[-6.21, -0.29]$	0.03
WhEng	-12.52	$[-24.22, -0.86]$	0.04
FillEng	-6.90	$[-19.47, 5.65]$	0.28
TopEng	7.30	$[-2.02, 16.65]$	0.13
TopGer	1.09	$[0.49, 1.69]$	~ 0
WhGer	-1.70	$[-4.16, 0.75]$	0.18
TopEng $\times$ FillEng	-8.47	$[-2.02, 16.65]$	0.11
SAEng $\times$ WhGer	4.43	$[0.85, 8.04]$	0.02
FillEng $\times$ WhEng	15.35	$[1.95, 28.78]$	0.03

Table 11: Model for German sentences with sentence adverbial as response variable with interactions. Green means significant (p < 0.05), red is not significant.

In Table 11 we observe the model with the best fit from the stepwise model selection.

Notice that all variables apart from the German filler sentences are present in the model, which may indicate that the model is overfitted to the observed data. Observe that the interaction terms have a positive effect on the model when the proportions of FillEng and WhEng are increasing, while English sentences as explanatory variables otherwise have a negative effect on the response when their proportion is increasing. This indicates that being proficient in English both increase and decrease the proportion correct answers for SAGer. We see no obvious pattern for the predicted proportion of SAGer for the model in Table 11, and we remake it. Eliminating the explanatory variables not present in the best fitted model for the additive model in Table 6. Using the stepwise model selection method to find the best fit, we get the model,

SAGer	Coefficient	C.I.	p-value
β_0	0.16	$[-1.76, 2.08]$	
SAEng	-3.20	$[-5.96, -0.45]$	0.02
WhEng	0.77	$[-0.08, 1.62]$	0.08
WhGer	-1.78	$[-4.06, 0.50]$	0.13
TopGer	1.08	$[0.49, 1.67]$	~ 0
SAEng $\times$ WhGer	4.53	$[1.18, 7.91]$	0.01

Table 12: Model for German sentences with sentence adverbial as response variable with interactions. Green means significant ($p < 0.05$), red is not significant, cyan is close to significant ($p \in (0.05, 1.00)$.

With the stepwise selection method on the model from Table 6 we get the model in Table 12. The interaction between German question sentences (WhGer) and SAEng is significant and have a positive effect on the response SAGer when the multiplied proportion value of the interaction SAEng $\times$ WhGer is increasing. Both of them separate have a negative effect on the response SAGer for increasing proportion correct answers, The linear predictor is,

$$\eta_i = 0.16 - 3.20 x_{SAEng} + 0.77 x_{WhEng} - 1.78 x_{WhGer} + 1.08 x_{TopGer}$$
$$+ 4.53 x_{SAEng} x_{WhGer} \tag{24}$$

and the corresponding plot
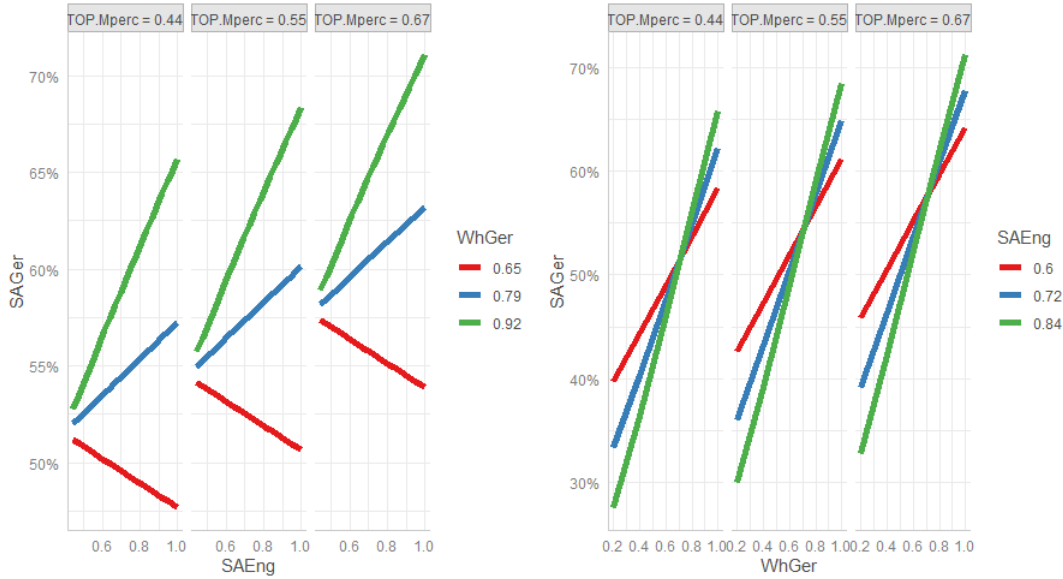
Figure 21: Model for SAGer as response variable with interactions SAEng×WhGer. Blue lines are estimated average for WhGer (left) and SAEng (right). Green lines are average proportion + 1 C.I.. Red lines are average proportion - 1 C.I.

The plots in Figure 21 shows two representations of the model in Eq. 24, for increasing value of proportion correct TopGer shown in the top of every plot. We observe in the left plot the effect of our interaction SAEng × WhGer. For high proportion correct answers WhGer we notice that for increasing value of the proportion correct answers for SAEng we also get an increasing in proportion correct answer for SAGer. When proportion for WhGer is below approximately 0.7 we see that increasing proportion of SAEng leads to a decrease in proportion for SAGer. This indicates that the positive effect from the interaction term is smaller than the effect from the significant explanatory variable SAEng.

Reason for this is the negative effect from SAEng when the proportion is increasing, which is observable in the left plot. Plot on the right side confirms this when we look at the green lines, which is the average proportion correct answer for SAEng with one confidence interval value added. They predict a low proportion correct answer of SAGer when WhGer has low proportion correct answers as well, similar to the red lines on the left side. We observe that when proportion of WhGer is increasing, the model rapidly predicts increasing correct answers for SAGer, which we observe for the green lines on the left side.

5.6   Discussions of RQ4 and RQ5

In this section the results are analysed in context of the research questions RQ4 and RQ5 from Chapter 1.1. We want to use the analysis for the models with response variable German sentences with topicalization and sentence adverbial to determine if there is evidence of L2 transfer. In the analysis we assume that transfer happens when participants score high on proportion correct answers for English sentences as explanatory variables that have a negative effect on a proportion correct answers German sentences as response variable. With high proportion correct answers of English sentences we want to see if high proficiency predict more or less L2 transfer to L3. That means we are looking for evidence that high proportion correct answers of English sentences leads to high or low proportion of German sentences correct. This is the reason we are focusing on the models with English sentences as explanatory variables for this research question.

In Table 1 we have models with one explanatory variable for response variable TopGer. We observe that the coefficient for TopEng as explanatory variable is negative. This indicates that an increase of proportion correct judged TopEng has a negative effect on the proportion judged TopGer, which means that the proportion decreases. We assume

that this is evidence that more proficiency in L2 leads to the participant judging L3 German sentences worse. There is a reason to assume an L2 transfer here, meaning that the participant possibly judging L3 sentences using the L2 structure that he or she got more experience with. The remaining models all have positive coefficient, thus an increase of the proportion correct explanatory variables leads to an increase of the proportion for the response. In Table 2 the response variable is German sentences with sentence adverbial (SAGer), and we have all positive coefficients and significant explanatory variables for all proportion correct English sentences. We understand this as an increase of the value for proportion correct answers of singular English sentences as explanatory variables have a positive effect on the response SAGer, and increases the proportion correct answers. If we consider RQ3 and look into the expectation from Dahl (2020) that there are more transfer for sentences with topicalization than for sentence adverbial, then we see this when the explanatory variable is TopEng. Significance of the explanatory variables also indicate that the models with SAGer as response are better than those with TopGer. Eq. 13 and 14 both indicate that participants with 4 years of German (YOG) judge the response variables TopGer and SAGer better than those with 5 YOG. Consider RQ5, and this can indicate that higher L3 proficiency, which we in this case assume participants with 5 YOG have, do increase L2 transfer slightly. The distribution of the data grouped on YOG is discussed more in Chapter 3.2. For the GLM, the difference is not big enough to say this is evidence.

We increased the explanatory variables for the model and found the best fit in Table 3, and we observe that TopEng influence the response variable TopGer negatively, meaning that an increase in the proportion correct answers TopEng decrease proportion correct answers for TopGer. Even if TopEng has a negative effect on the proportion for the response, the positive effect from English question sentences (WhEng) is bigger. We also assume that the participants in average get higher proportion correct answers for WhEng than for TopEng, making it a more significant explanatory variable for the model than

TopEng is, which is also shown to be true in Table 3. We assume no evidence of L2 transfer in this model, but both Table 1 and 3 show that there is a negative effect from TopEng on response TopGer. Further investigation in Table 7 show evidence of an interaction between TopEng and WhEng. The analysis also indicate that there is evidence of transfer when introducing an interaction term. We see in Figure 17 that when a participant has a high proportion correct answers for both TopEng and WhEng, it leads to a decrease in proportion correct answers for response TopGer. Considering RQ4 we see evidence of L2 transfer, reason being that high L2 proficiency leads to a participant making more incorrect judgements of German sentences. We see that for a low proportion of WhEng and TopEng, participants have a low proportion correct TopGer, which can be evidence that a participant that is worse at recognizing correct sentences may just struggle with sentences structure and learning L2 and L3 overall. The increased interaction model in Table 8 confirmed the evidence of L2 transfer too.

With response variable German sentences with sentence adverbial (SAGer) we see in Table 5.4 that an increase in proportion correct answers for TopEng got a negative effect and decreases proportion correct SAGer. It is not significant, and with the positive effect from English sentences with sentence adverbial (SAEng) and English question sentences (WhEng) being bigger and significant we assume that with higher L2 proficiency there is less transfer. We also observe in Figure 15 an increase in proportion correct answers for SAGer when all other explanatory variables are increasing. The models with interaction terms in Table 10 predicts a negative effect from the interaction term with SAEng and TopEng with some significance. That is, if a participant is good at judge correctly on these sentences we see a decrease in proportion correct SAGer, which indicate L2 transfer. The analysis however pointing towards less transfer, because of the lack of significance and the potential overfitting for the model. Figure **??** can be misleading, as it predicts both increase and decrease in value of the proportion correct answers for SAGer depending on what explanatory variable we look at in the

plots. So it is unclear if L2 proficiency predict more or less transfer for this model.

When we increase the models including all explanatory variables, we see in Table 4 that German sentences have a positive effect on the response, as well as being significant for the model. FillEng having a negative effect does assume that high proficiency may lead to L2 transfer. Consider RQ5, and we look for how high L3 proficiency effects transfer. High L3 proficiency means that we look at the effect on proportion for TopGer when the proportion correct answers of German sentences as explanatory variables is increasing. Table 4 indicates that high L3 proficiency leads to less transfer from L2. Reasons being the significant explanatory variables and higher coefficienct for German sentences. We also notice in Figure 14 that proportion of TopGer is increasing for an increase in proportion correct answers of other German variables. Interactions introduced in Table 9 indicates much of the same. We notice that participants that are correctly judging WhGer and SAGER with high proportion, showing high L3 proficiency, also get high proportion correct answers of TopGer. For less proficiency in L2, which we assume is when proportion correct answer for FillEng is low, we observe an increase in proportion correct answers for TopGer. The effect of high L3 proficiency makes L2 transfer less pronounced. Reason for this is the unchanged proportion of TopGer when participants get a high proportion of FillEng correct.

With SAGer as response we observe in Table 6 that a high proportion for all sentences predict high proportion correct answers for SAGer. This indicates that with high L3 proficiency an L2 transfer is less pronounced, but with every explanatory variable having a positive effect on the proportion of SAGer this also indicate that high L2 proficiency makes transfer less pronounced. With the interaction in Table 12 we observe that both SAEng and WhGer are effecting the proportion correct answers for SAGer negatively individually. The interaction between them has a positive effect, but SAEng has a higher coefficient as explanatory variable. This means that with high L2 proficiency we predict

L2 transfer. With high L3 proficiency and high L2 proficiency we see for this model that the transfer is less pronounced. This is observed in Figure 21 where high proportion correct answers for both of the interaction terms also increase the proportion correct answers for SAGer. We also observe when proportion for WhGer is low, the proportion correct answers for SAGer decreases for increasing proportion correct answers of SAEng. This is evidence of less L2 transfer with high L3 proficiency.

6   Conclusion

The goal of this thesis is to see if GLM and proportion of correct answers can be used to analyze AJT-data, which is data on the form of a 1-4 Likert scale. The proportions were found by assuming a three-trial binomial method, that every participant gets to make a judgement for the same sentence three times. To analyze we tried to solve the research questions from a linguistic research paper in light of the statistical models that are created using GLM. We were looking for evidence of potential transfer of verb placement from English to German, given that a participant got more experience learning English. We observed that some models predicted that participants do worse judging German sentences when they do well judging English sentences in the case when only proportion correct answers for English sentences were used as explanatory variables. For an increased model, this effect got less likely and we introduced models with interaction term to investigate further. This did show that interaction between English and German sentences lead to a participant judging German sentences worse. We also found out that high proficiency in German do make transfer from English less likely. In Chapter 3.2 we found that there can be evidence of English language transfer to Norwegian learners of German by using the proportion chosen Likert score for each sentence type divided into the participants respective years learning German. We observed that there is little

difference between learning German for four or five years, while learning English there was little difference between any of the groups.

In this thesis we have only skimmed the surface, and we recommend to increase the tools for model fit and explore how Mixed Models can be included to analyze these data. In this thesis we have shown that these methods can give new interpretation opportunities for linguistic research.

References

Bardel, C., & Falk, Y. (2007). The role of the second language in a third language acquisition: the case of germanic syntax. *Second Language Research*, 459-484.

Dahl, A., Listhaug, K. F., & Busterud, G. (2020). *A case study of The role of L1 Norwegian and L2 English in the acquisition of verb placement in L3 German. [Unpublished Manuscript]*.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models*. Taylor Francis Group, 3. ed.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression models, methods and applications*. Springer-Verlag Berlin Heidelberg, 1. ed.

Jackson, F. H., & Kaplan, M. A. (1999). Lessons learned from fifty years of theory and practice in government language teaching. *Round Table on Languages and Linguistics*, 71-87.

Robson, C., & McCartan, K. (2016). *Real world research*. John Wiley and Sons Ltd, 4. ed.

Zhang, Z. (2016). Variable selection with stepwiseand best subset approaches. *Annals of translational medicine*, *4*(7), 136. doi: https://doi.org/10.21037/atm.2016.03.35

A Coefficients and p-values Figure 12

	x_{ij}	Coefficient	p-value
1 YOG	a)	-0.38	0.30
	b)	0.02	0.95
	c)	0.09	0.66
	d)	0.30	0.04
2 YOG	a)	0.16	0.36
	b)	0.17	0.18
	c)	0.13	0.56
	d)	0.07	0.68
4 YOG	a)	0.15	0.29
	b)	0.11	0.35
	c)	0.20	0.15
	d)	0.32	0.00
5 YOG	a)	-0.30	0.10
	b)	-0.19	0.27
	c)	0.23	0.27
	d)	0.40	0.03

Table 13: Coefficients for explanatory variables and significance level (p < 0.05) for Figure 12

.

B stepAIC tables

Step	AIC
1	756.23
2	755.05
3	754.57

Table 14: Stepwise selection method for German sentences with topicalization as response for Table 3

Step	AIC
1	691.56
2	689.58
3	688.22
4	687.86

Table 15: Stepwise selection method for German sentences with topicalization as response for Table 4

Step	AIC
1	749.39
2	747.36
3	745.38
4	743.57
5	741.39
6	740.39

Table 16: Stepwise selection method for German sentences with topicalization as response for Table 8

Step	AIC
1	690.05
5	682.18
10	673.62
16	666.24

Table 17: Stepwise selection method for German sentences with topicalization as response for Table 9

Step	AIC
1	697.26
2	696.54

Table 18: Stepwise selection method for German sentences with sentence adverbial as response for Table 5

Step	AIC
1	648.32
2	646.70
3	645.21
4	643.90

Table 19: Stepwise selection method for German sentences with sentence adverbial as response for Table 6

Step	AIC
1	702.84
2	700.91
3	699.22
4	697.80
5	696.76

Table 20: Stepwise selection method for German sentences with sentence adverbial as response for Table 10

Step	AIC
1	648.29
2	646.29
3	644.36
4	642.45
5	640.58
6	638.85

Table 21: Stepwise selection method for German sentences with sentence adverbial as response for Table 12