Stockholms universitet

Acquisition of grammatical gender and number agreement in Italian as a second language - A statistical analysis

Magnus Gudmundson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:2
Matematisk statistik
Maj 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2014:2**
http://www.math.su.se

Acquisition of grammatical gender and number agreement in Italian as a second language - A statistical analysis

Magnus Gudmundson[*]

Oktober 2014

Abstract

In this bachelor thesis we aim to achieve knowledge of which factors that can explain the difficulties with correct use of number and gender agreement while learning Italian as a second language. Previous studies has looked at different factors, such as measures of lexical diversity and the availability and reliability measures, in an univariable way. Aim of this thesis is to investigate these factors in a joint statistical model to answer a given number of hypothesis directly related to the considered factors. The collected data are the binary outcomes of correct or incorrect use of number and gender agreement in transcribed interviews of Swedish students studying Italian at the university. Due to the way data has been collected it possess an unbalanced nested structure which in combination with the binary outcome suggests a rather complex hierarchical modelling approach. But the question whether it is even possible to fit complex model using maximum likelihood estimation arises from the fact that the majority of outcomes are cases of correct use why we most likely will face the numerical problem of separation. In the analysis we thereby adopt the strategy of fitting a less complex base model to investigate the limits of a maximum likelihood approach and to get an idea of which covariates to include in a more complex model. The results of the final base model in terms of the hypothesis are presented but the interpretation of the results should be cautious due to the violation of the model assumption of independent samples. From the fitting of the base model we can conclude that we are not able to proceed with a extended analysis using maximum likelihood estimation. Ideas and suggestions for further analysis based on Bayesian inference are discussed but not explicitly presented.

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:nohands@gmail.com . Supervisor: Michael Höhle.

Acknowledgements

I would like thank my supervisor Michael Höhle for great encouragement and advice, and for introducing me to the concept of time. Thank you Anna Gudmundson for letting me take part of your work.

Contents

1 Introduction

Difficulties in language acquisition differ between learner and learner, but some difficulties are due to the characteristics of the specific language. All languages have some degree of an intricate unique semantic structure in terms of irregularities of the way words are connected to form, what one intuitively would like to call, well formed sentences. These irregularities could be considered problematic for any learner and even in some cases problematic for the experienced user. In this bachelor thesis we will conduct

a statistical analysis by taking a closer look at the the specific problem of number and gender agreement in learning Italian as a second language. This field of studies in language acquisition is relatively unexplored in terms of thorough statistical methods why a statistical analysis is of special interest. In writing this thesis we used knitr (Xie et al., 2013) for dynamic report generation in R (R Core Team, 2013) with output language LaTeX.
We will in this section start by in section 1.1 taking a look at some basic Italian grammar and in section 1.2 we explain the two concepts types and tokens. We end by, in section 1.3, introducing the theory of the Competition Model.

1.1 Grammatical gender and number agreement in Italian

To be able to grasp the notion of number and gender agreement in Italian we give here a short repetition of grammatical fundamentals such as noun, adjective, number and gender. Those who are familiar with the basics of Italian grammar can skip this part.

1.1.1 Some grammar

A *noun* is a class of words that normally represent an object, abstract or concrete, such as ”car”, ”city, ”mother”, ”thought” or ”math department”. The *adjective* on the other hand describes the specific object by naming some characteristics such as ”blue”, ”tall”, ”hungry” or ”impossible”. With the two classes together we get things like ”tall mother”, ”blue car” and ”impossible thoughts”.
The last example where the noun ending ”s” was used gave us information about that it is not just one impossible thought but several of them. This introduces the concept of *number* that is whether the noun denotes one (*singular*) or several (*plural*) objects.
As a final step we need the *gender*. In some languages, including Italian, the nouns are classified as being either *masculine* or *feminine*. Just as information about number, information about gender is concealed within the noun and normally in terms of a specific ending. The difficulties of learning how to interpret and use this information while learning Italian as a second language is the main concern of this thesis.

1.1.2 Italian grammar

In this thesis the data consists of transcribed interviews of students studying Italian. The main interest is in trying to find factors which help explain the difficulties of learning how to correctly use number and gender when combining an adjective with a certain noun.
Let’s take a look at some basic examples to get an idea of the problem at hand. The sentences ”dear brother”, ”dear sister”, ”dear brothers”, ”dear

sisters” translates respectively into Italian as ”caro fratello”, ”cara sorella”, ”cari fratelli” and ”care sorelle”. In ”caro fratello” the noun ”fratello” means brother and the adjective ”caro” translates as ”dear”. Obviously ”fratello” is being classified as masculine and, here as in the other examples, we can see that the adjective ending in ”caro” inherits the noun ending which in this case is ”o” and thereby also carrying the information *masculine singular*. When the information in the noun and adjective agree, that is being used correctly, a speaker achieve number and gender agreement. In ”care sorelle” translated as ”dear sisters” where the noun carries the information *feminine plural* the ending ”e” indicates that this might be the right interpretation. The above examples are pretty straight forward in the sense that the gender of the objects under consideration is obvious. In general this is not the case and no strict rules are available to guide the learner. For example in ”lavoro dificile” translated as ”hard work” the noun ”lavoro” indicates masculine singular but the adjective ”dificile” seems to carry the information feminine plural. Of course this is not a case of non agreement. In fact the ending ”e” in ”dificile” is the same for both feminine and masculine in singular. These types of agreements cause troubles for the new learner and in could in some sense be considered as general problems.

1.2 Types and tokens

Two other concepts, types and tokens, will occur frequently. As an explanation consider the set $\{0, 1, 1, 0, 1, 3\}$. In this set we have the three types 0, 1 and 3 and six tokens. That is the digits represent tokens and the types represent the specific number of the digit. Another example, consider the sentence ”I studied and I studied”. Here ”I”, ”studied” and ”and” are the three types but we have a total of five tokens.

1.3 The Competition Model (CM)

The main questions and hypothesis considered in this thesis, presented in the next section 1.4, have their origin in a specific theoretical background. Here a short presentation is given of the theory to elucidate and motivate the questions asked. For a comprehensive presentation see MacWhinney (2005). The theory of the Competition Model (CM) is partly built upon the idea that language can be described as a distributional system where different forms and functions are mapped to one another in terms of a probabilistic network. A function can be considered as the meaning of a specific expression where the expression is the actual form. Different forms/expressions may be used to express the same function/meaning and different functions may be used to interpret a specific form. When there are multiple choices of forms and functions we say that these ”compete” with one another in terms of our probabilistic network.

So in the spirit of CM language acquisition can be thought of as building a probabilistic system where the building process partly consists of updating our network given a linguistic input. As a consequence frequencies and regularities of such input are of importance. These frequencies and regularities the CM linguistics tries to summarize as frequency measures such as availability and reliability (see section 2.1.4).
The idea of competition in terms of probabilistic network have its origin within the area of psycho-linguistics and cognitive science.

1.4 Aims of the analysis

In previous studies on number and gender agreement in learning Italian as second language the distributional characteristics under consideration have been investigated using mainly descriptive methods. In those studies where statistical methods have been used the factors of interest, such as availability and reliability, have been considered in a univariable way. The main aim of this thesis is to consider all factors in a joint statistical model and under this model answer the following four hypothesis.

- Singular agreement will have higher rates of correct use than plural and masculine agreement will have higher rates of correct use than feminine. This is motivated by the fact that singular and masculine has higher frequency rates in the language input.

- Higher values of the frequency measures availability and reliability, defined in section 2.1.4, will have a positive effect on correct use in terms of number and gender agreement.

- That learners of Italian with a higher linguistic level have higher rates of correct use. The VOCD measure introduced in section 2.1.3 will be used as a measure of linguistic level.

- That learners of Italian as a second language get better on using number and gender agreement over time.

2 Data of the InterIta corpus

The data used in this thesis has been extracted from the corpus InterIta created at the Department of French, Italian and Classical languages at Stockholm University, see Bardel (2004).
InterIta is composed of 71 transcribed recorded interviews in which Swedish students, studying Italian at different levels at Stockholm University, have been asked to talk in Italian about things such as their family, Italy and future projects. Thus the interviews are conducted as natural dialogues.

All 25 students who signed up for the project did so without any sort of compensation and participated only out of interest.

Some of the individuals has been interviewed just once while others has been interviewed five or six times, see table 1 for a summary of interview frequencies. The dates of the recordings have been chosen in terms of the availability of the participants and hence does not follow a specific pattern. Each interview have a recording length of approximately 25 minutes.

no interviews	1	2	3	4	5	6
no stud	5	7	4	6	2	1

Table 1: Number of students in each interview frequency

2.1 The variables

The data under consideration in this thesis is the extraction of all noun adjective combinations extracted from the interviews in InterIta . This was done in Gudmundson (2012). For examples on noun adjectives combinations see section 1.1. The process of extracting was done by first letting a computer program run through the entire data set and tag each word with it's grammatical characteristics. In the next part of the extraction a Perl Script was created to extract the noun adjective combinations and organize them accordingly. As a last step all combinations were run trough manually and classified as either correct or incorrect use in terms of number and gender agreement.

The resulting structure of data is as as follows. We have a total of 25 students which have been interviewed a different number of times as presented in table 1. From each of these 71 interviews all noun adjective combinations where extracted. The number of extracted combinations differ of course from interview to interview but the total number sums up to 3177. What we end up with is a unbalanced nested structure with noun adjective combinations nested within interviews and interviews nested within students.

The available data on the students, the interviews and the noun adjective combinations leads to the definition of the following set of variables which we will use as the base of the statistical analysis.

2.1.1 Outcome

The outcome, that is whether the speaker manages to combine the noun and the adjective in a correct way in terms of number and gender, we define as the binary variable *corr.use*. So *corr.use* takes value 1 if correct use and 0 otherwise.

To emphasize the nested structure we consider it here in terms of *corr.use* where we enumerate a specific outcome as $y_{j,k,l}$, $j = \{1, .., 25\}$, $k = \{1, .., n_j\}$,

$l = \{1, ..., n_{j,k}\}$, where j refers to the level of students, k refers to the level of interviews and l refers to the level of noun adjective combinations.
In the following sections we define the rest of the variables following the hierarchical structure as we move our way up the three levels.

2.1.2 Student based variables

We start by introducing the variables denoting the characteristics of the students. The name (not the real name), and sex of the 25 student we define as the nominal variable *name* and the binary variable *sex*.

The students age in years at the time of the first recording occasion will be represented by the ordinal variable *age.first* with possible values in $\mathbb{Z}_+$.
Each student has spent a different amount of time measured in months in Italy at the time first interview this time we define as the continuous variable *ital.time* taking values in $\mathbb{R}_+$.
Number of semesters spent studying Italian at some university we define as the ordinal variable *univ.time* with values in $\{0, 1, 2, 3\}$. An other factor of interest is the number of Roman based languages the students to some degree can speak. Roman based languages are similar in structure and use and knowing one will probably make it easier to learn a new one. To the family of Roman languages counts for example French, Spanish and Portuguese. The number of Roman languages the student know as a second language at the time of the first interview we define as the ordinal variable *roman* with values in $\{0, 1, 2, 3\}$.

2.1.3 Interview based variables

The interview number in terms of order we define as the ordinal variable *int.numb* with possible values in $int.numb \in \{1, ..., 6\}$
The date of the recorded interview will be represented by the variable *rec.date* with values such as "2001-09-21". Using these dates we define one additional variable which gives the time, measured in months, since the first interview for a specific student and interview occasion. In notation we will represent this variable as *months.rec*.

VOCD (MacWhinney, 2000) is a measure of lexical diversity which can be described as the range and diversity of the vocabulary of a specific text. This metric will be used as information about the level of speech of a transcribed interview where higher values indicates higher levels. A more basic measure of the same kind is the Type Token Ratio where the number of types is divided by the number of tokens. The Type Token Ratio exemplifies the main problem with these kinds of measures namely that as soon as the length of a text increases, in other words the number of tokens increases, the ratio drops in value. This problem is supposed to be dealt with by using the VOCD measure which under the right circumstances is supposed to be more or less independent of the total number of tokens in the text under consideration. VOCD we define as the continuous variable *vocd* with value domain $\mathbb{R}_+$

The total number of spoken tokens during the approximately 25 min long interview we define as the ordinal variable *token.freq* with value domain $\mathbb{Z}_+$.

2.1.4 Adjective noun combination based variables

Finally we consider those variables with information on the noun adjective combinations.
The ending of the noun and the ending of the adjective in the combination we define respectively as the nominal variables *noun.end* and *adj.end* where

$$noun.end = \begin{cases} a & \text{if ends in ``a''} \\ A & \text{if ends in ``à''} \\ c & \text{if ends in ``consonant''} \\ e & \text{if ends in ``e''} \\ i & \text{if ends in ``i''} \\ o & \text{in ends in ``o''} \end{cases}$$

$$adj.end = \begin{cases} a & \text{if ends in a} \\ c & \text{if ends in ``consonant''} \\ e & \text{if ends in ``e''} \\ i & \text{if ends in ``i''} \\ o & \text{in ends in ``o''} \end{cases}$$

When the noun and the adjective in the combination have the same ending, as in *care sorelle* (dear sisters) where both end in *e*, there is assonance. One can say that the noun and adjective more or less rhymes under these circumstances. Assonance we define as the binary variable *asso*.

Next we take a look at the measures availability and reliability. The purpose of the two is to describe how common or unique a specific class of nouns is in relation to other classes within a specific language. The classification is defined by noun ending, gender and number. As a representation of the Italian language the corpus LIP created by De Mauro et al. (1993) have been used. It's a summary of numerous discussions taking place in four different Italian cities over a two year period. In other words transcribed discussions of Italians speaking Italian. The frequencies of the noun classes in LIP are then used as representations of the actual frequencies in the Italian language. By these frequencies one calculate the availability as below where the rather sloppy notation $\{nounending, gender, number\}$ is to be interpreted as the class of all nouns with a specific noun ending, specific gender and a specific number.

$$\frac{\#\{nounending, gender, number\}}{\#\{gender, number\}}$$

That is how common is this noun ending given the specific gender and number. Reliability is calculated as

$$\frac{\#\{nounending, gender, number\}}{\#\{nounending\}}.$$

Here the ratio is instead being expressed with the total number of occurrences of the specific noun ending in the denominator. This can be interpreted as how common the gender and number combination is given the the specific noun ending. Intuitively these measures might explain difficulties with the learning process of number and gender agreement in terms of how common one ending is or how unique it is. So the availability and the reliability of the noun in a specific noun adjective combination we define as the two continuous variables *noun.reliab* and *noun.avail* taking there values on the interval $[0, 1]$.
Another measure of interest is the validity which is defined as the product of the availability and the reliability, that is the interaction between the two. The validity we define as the continuous variable *noun.valid* which of course also takes its value on the interval $[0, 1]$.

2.1.5 Variable summary

The variables described in the previous text we summarize in the following table.

Level	**Abbreviation**	**Description**	**Type**	**Values**
Student				
	name	Student Name	Nominal	$\{Alice, Ulla, ...\}$
	age.first	Student Age	Ordinal	$\mathbb{Z}_+$
	female.male	Student Sex	Binary	$\{female, male\}$
	ital.time	Time in Italy	Continuous	$\mathbb{R}_+$
	univ.time	Semesters	Ordinal	$\{0, ..., 3\}$
	Roman	Languages Spoken	Ordinal	$\{0, ..., 3\}$
Interview				
	int.numb	Interview Number	Ordinal	$\{1, ..., 6\}$
	rec.date	Recording Date	Ordinal	{2001-09-21,...}
	vocd	VOCD	Continuous	$\mathbb{R}_+$
	token.freq	Token Frequency	Ordinal	$\mathbb{Z}_+$
	months.rec	Months since first Interview	Continuous	$\mathbb{R}_+$
Noun				
Adjective	*numb*	Number	Binary	{sing, plur}
Combination	*gend*	Gender	Binary	{fem, masc}
	noun.end	Noun Ending	Nominal	{a, A, c, e,i,o}
	adj.end	Adjective Ending	Nominal	{a, c, e, i, o}
	asso	Assonance	Binary	{0,1}
	noun.reliab	Noun Reliability	Continuous	$[0, 1]$
	noun.avail	Noun Availability	Continuous	$[0, 1]$
	noun.valid	Noun Validity	Continuous	$[0, 1]$
Outcome				
	corr.use	Correct Incorrect use	Binary	{0,1}

3 Descriptive analysis

In this section we will examine the distributional characteristics of each variable included in the data set. We will try when possible to implement the nested structure into the descriptives but most of the variables characteristics will be examined just on its specific level. To get an overview of the available data and variables described in the previous section, table 2 shows the first five rows in the of the data set.

	name	age.first	sex	int.numb	rec.date	vocd	token.freq
1	Alice	19	f	1	2001-09-21	33.220	1070
2	Alice	19	f	1	2001-09-21	33.220	1070
3	Alice	19	f	1	2001-09-21	33.220	1070
4	Alice	19	f	1	2001-09-21	33.220	1070
5	Alice	19	f	1	2001-09-21	33.220	1070

	numb	gend	noun.end	adj.end	asso	noun.reliab	noun.avail
1	sg	f	a	c	0	0.946	0.642
2	sg	f	a	a	1	0.946	0.642
3	sg	f	a	a	1	0.946	0.642
4	sg	f	a	a	1	0.946	0.642
5	sg	m	c	c	0	0.897	0.016

	noun.valid	corr.use	error	ital.time	univ.time	roman
1	0.608	1	c	0.5	1	1
2	0.608	1	c	0.5	1	1
3	0.608	1	c	0.5	1	1
4	0.608	1	c	0.5	1	1
5	0.015	1	c	0.5	1	1

Table 2: First five rows of the data set.

We now proceed by examine the data, variable by variable and just as in section 2 following the levels of the hierarchical structure.

3.1 Student descriptives

Figure 1 contains a barplot of the ages of students at their first recording occasion and their sex. The median age is 25 years. We also notice that only 5 out of 25 students are males.

Looking at figure 2 we have *corr.use* from the first interviews plotted against *ital.time* the time spent in Italy. This type of plot will be used frequently through out the descriptive analysis. In this plot the size of

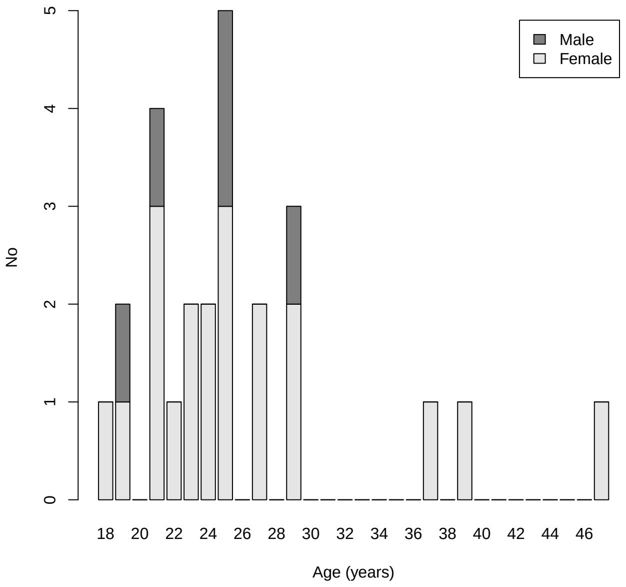

Figure 1: Age and Sex distribution of the 25 interviewed persons.

the circles are proportional to the frequency of observations in that specific point. The LOESS curve, Cleveland (1979), is a non parametric regression method which here is used strictly as a descriptive tool as an indicator of trends. Now looking at figure 2 once again we see an indication of that those students who spent more time in Italy performed better at the time of the first interview.

The number of Roman languages the student know as a second language, which we see in table 3, seems to have a negative effect on the rate of correct use. This might seem counter intuitive but from the view that being able to speak a number of similar languages makes things more ambiguous, it's not.

0	1	2	3
0.94	0.90	0.89	0.84

Table 3: The mean rate of correct use given number of Roman languages spoken by the student.

In figure 3 the mean rate of *corr.use* per interview is plotted against the time between each interview occasion for each participant with more than one interview. The plot is divided in to four separate plots to easier get an overview of the respective students. The lack of incorrect uses is apparent, that is the majority of the recorded students have been performing to well.

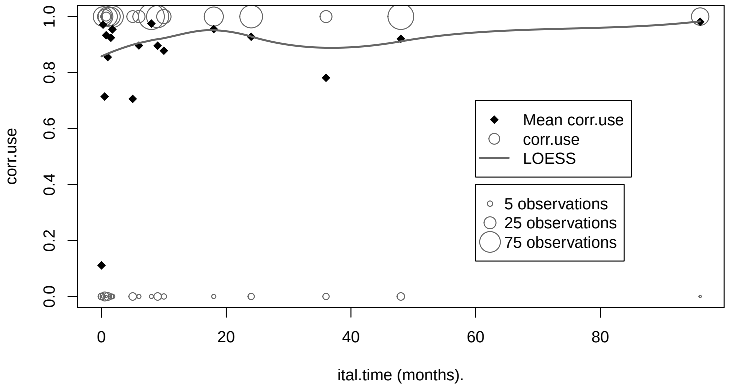

Figure 2: Descriptive illustration of the association between time spent in Italy at the time of first interview an the outcome. Observations are shown as circles proportional the frequency. Furthermore, the average per time unit and a LOESS curve are shown.

Only 292 of the 3177 outcomes are incorrect which might be a problem for the construction of a joint model. One can see that there is a huge variation in individuals in terms of initial values and their development over time. Few students shows indication of improvement and no overall trend is apparent. This might be explained by the fact that most students seem to be already on really high linguistic level. Nonetheless they still have problems of getting the noun and gender agreement right and what factors that are of significance in explaining those problems is of interest.
There is one obvious outlier with an initial value of approximately 0.1 in the lower left plot. A closer examination of the student reveals that the person does not fulfill certain criteria to be considered a suitable representative of the population under consideration. The student is removed from the data set and will not be considered in the further analysis. Before proceeding we make sure that the rest of the student sample fulfill the population criteria.

Those students with only one interview we can see in table 4. The values are extremely high but not extreme in relation to what we see in figure 3.

	Cecilia	David	Frank	Kristina	Sandra
Mean of *corr.use*	0.905	0.971	0.967	0.981	0.924

Table 4: Mean of correct use for students with just one interview.

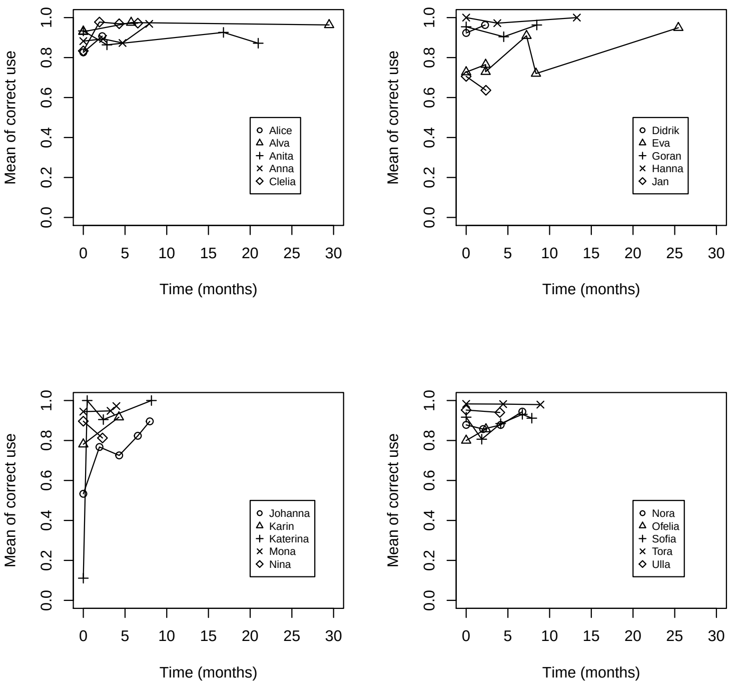

Figure 3: Time since first interview of those students with more than one interview occasion.

3.2 Interview descriptives

VOCD which we in section 2.1.3 introduced as a measure of lexical diversity with the purpose to serve as an representation of the level of speech. We would then expect that an increase in VOCD suggests a higher rate in correct use. If we look at figure 4 there is some indication of an increase of correct use with higher values of VOCD. For values larger than 80 the variance show signs of being stable compared with values below 80 where the variance is greater and more unstable.
Looking at the rate of correct use against token frequency figure 5 tells us than an increase in frequency means higher rates of correctness. The variance follows a similar pattern by decreasing with higher values. If we with higher levels of speech mean higher rates in correct use it seems like the number of uttered tokens during the 25 min interview might be a better measure.
The plot in figure 6 indicates some positive correlation between token frequency and VOCD.

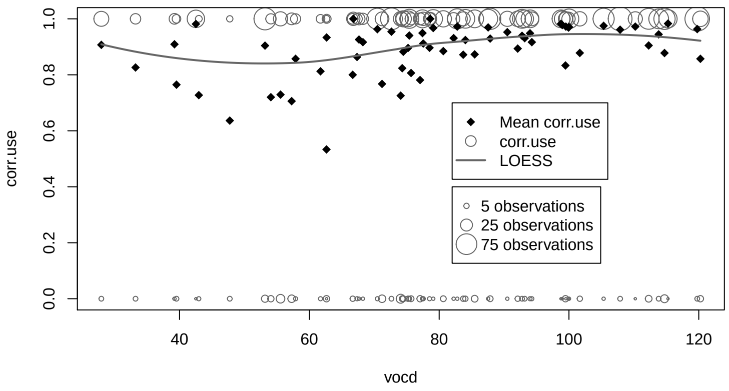

Figure 4: Correct use against VOCD measure.

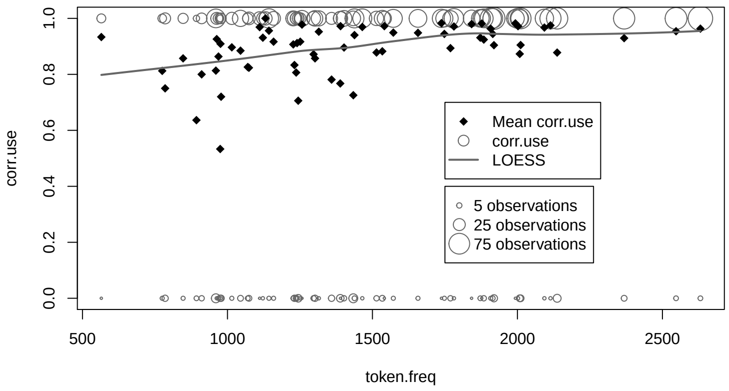

Figure 5: Correct use against token frequency.

3.3 Number and gender descriptives

Now let us follow the line of the token frequency as an indicator of level of speech. Splitting the data given the binary variables number, gender and assonance respectively we plot once again correct use against token frequency but this time given the binary variables. This will show us how the token frequency and the binary variables are confounded but more importantly it will show how they affect the mean rate of correctness within each interview and thereby reveal an eventual general pattern.
Looking at the token frequency given gender in figure 7 we see that given masculine the mean rate of correct use is overall higher than given feminine.

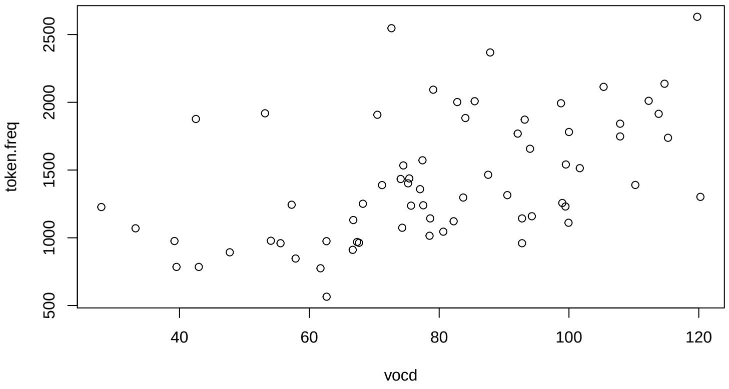

Figure 6: Token frequency against VOCD.

For higher frequencies feminine seems to stable around 0.9 and masculine around 0.96 indicating that feminine gender is the harder gender to master in terms number and gender agreement. The marginalized relation between gender and correct use presented in table 5 tell us as expected that the rate of correct use is higher in cases of masculine gender. A Chi-Square test generates a p-value of 0 indicating dependence on this general level not considering confounding factors. The Chi-Square test implemented in this way is the statistical method used in previous studies.

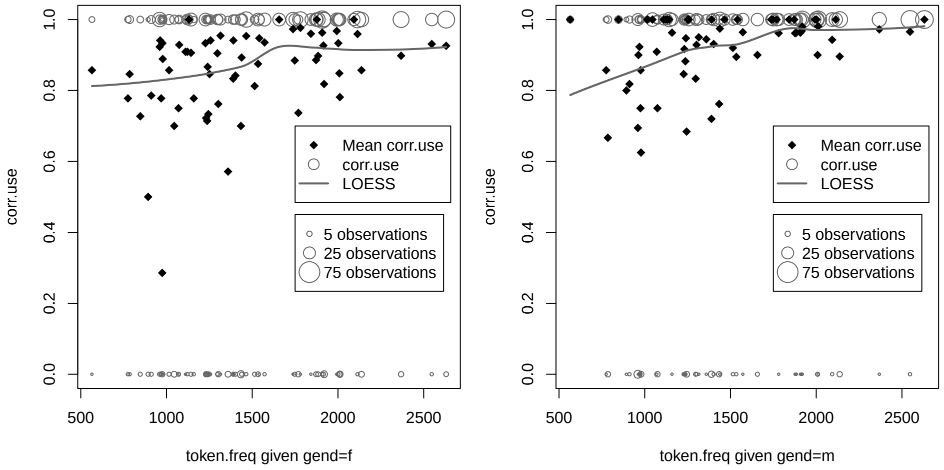

Figure 7: Correct use against the token frequency given feminine and masculine gender respectively.

The token frequency given number tells a similar story as in the case of

	Gender			
	f		m	
Correct	Percent	All	Percent	All
0	11.5	179	6.558	101
1	88.5	1377	93.442	1439

Table 5: 2 × 2 table over gender and correct use with column proportions.

gender. We see in figure 8 that token freq given singular seems easier to master than plural which overall have a lower rate of correct use. Variance is also greatly reduced given singular compared to given plural. In table 6 the results are again as expected with a higher rate of correct use given singular. The Chi-Square test generates a p-value of 0 telling us that there is dependence between number and correct use under the marginalized distribution. The previous results of gender and number are both consistent with hypothesis one stated in section 1.4.

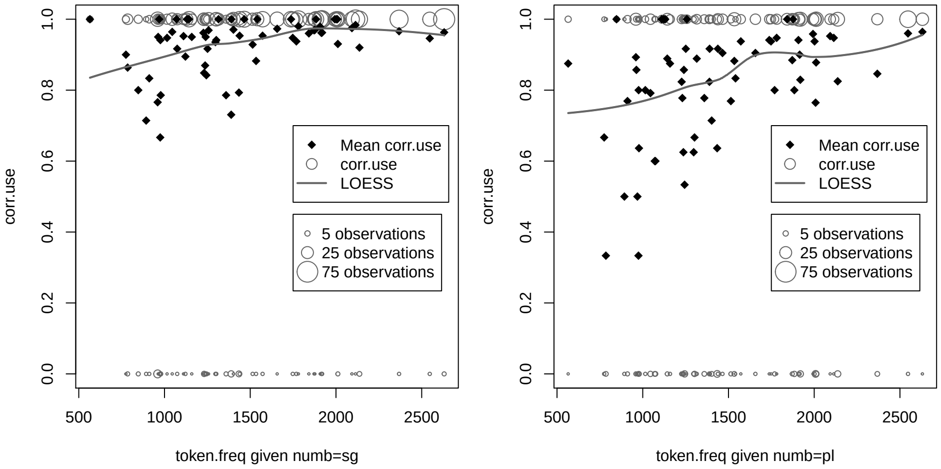

Figure 8: Correct use against the token frequency given singular and plural number respectively.

When there is assonance, that is whether the noun and adjective have the same ending, we intuitively would say simplifies the learning process. Figure 9 verify this assumption to some extent. There is an overall difference but for higher values of token frequency the partial distributions does not show much difference. One big difference we can see in that for values larger than 1500 there is almost no variance given there is assonance which overall have a smaller variance compared to when there is none. The 2 × 2 contingency table of the marginalized distribution in table 7 gives us with a p-value,

	Number			
	pl		sg	
Correct	Percent	All	Percent	All
0	14.61	168	5.755	112
1	85.39	982	94.245	1834

Table 6: 2 × 2 table over number and correct use with column proportions.

generated by the Chi-Square test, of 0.003 telling us there is dependence between correct use and assonance.

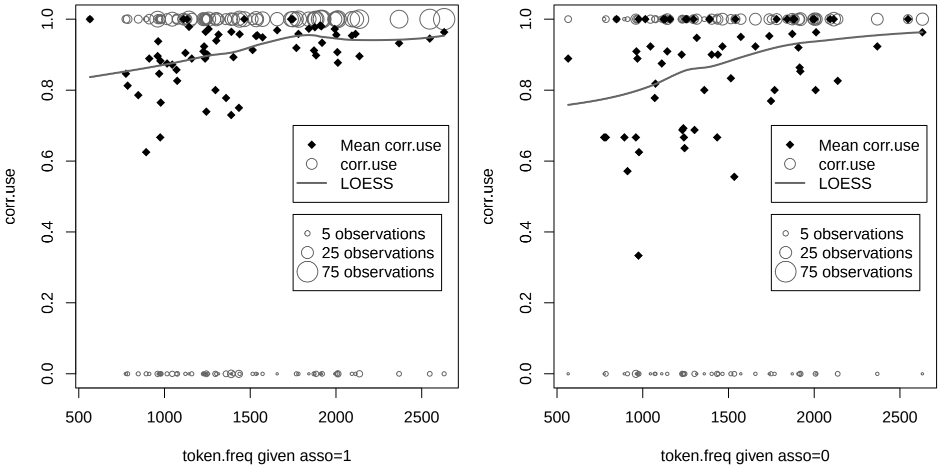

Figure 9: Correct use against the token frequency given assonance and not assonance respectively.

	Assonance			
	0		1	
Correct	Percent	All	Percent	All
0	11.43	109	7.983	171
1	88.57	845	92.017	1971

Table 7: 2 × 2 table over assonance and correct use with column proportions.

Recalling hypothesis two from section 1.4 we are interested in the ratio of correct use in relation to the frequency measures noun reliability and noun availability. In figure 10 correct use is plotted against the two. Noun reliability show signs of improvement for higher with an increase in value and is more or less linear in shape. This is consistent with the hypothesis which states that higher values of reliability means higher rates of correct use. Noun availability show no signs of linearity and the LOESS curve takes values outside the range of the response. We could have used a GAM curve with a binary response instead to avoid this problem.

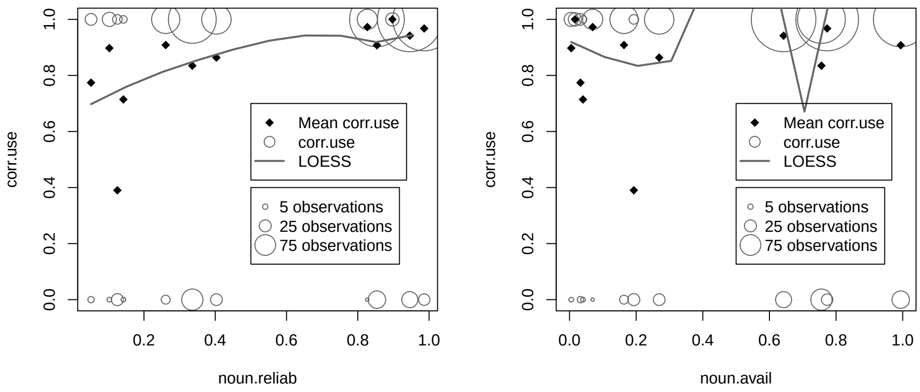

Figure 10: Correct use against noun-reliability and noun-availability respectively.

Something intuitively important when looking at number and gender agreement are the different endings of the nouns and the adjectives. From a CM perspective (see section 1.3) the frequencies of the different endings in the Italian should be if interest. No such frequencies are available in the data but we can consider the classification due to the endings them self. In table 8 we see that the different noun endings show signs of big differences in both frequency of use and proportion of successes. These differences indicate a dependence between the rate of correct use and the noun endings. We observe a similar relation of dependence between correct use and the adjective endings looking at table 9.

	Correct			
	0		1	
Noun	Percent	All	Percent	All
a	6.343	51	93.66	753
A	5.814	5	94.19	81
c	5.714	4	94.29	66
e	14.815	120	85.19	690
i	12.804	79	87.20	538
o	2.962	21	97.04	688

Table 8: 6 × 2 table over noun ending and and correct use with row proportions.

	Correct			
	0		1	
Adjective	Percent	All	Percent	All
a	8.9947	68	91.01	688
c	0.8547	1	99.15	116
e	7.9365	65	92.06	754
i	14.8668	106	85.13	607
o	5.7887	40	94.21	651

Table 9: 6 × 2 table over adjective ending and and correct use with row proportions.

4 Statistical modeling

As mentioned in section 3.1 there is an overall high ratio of the response variable ”correct use” being equal to one. As a consequence we will most likely face difficulties if aiming to fit a model with some degree of complexity due to problem of separation among the observations and hence non-convergence of the IRLS (iterated re-weighted least square) algorithm. Because of these difficulties our modeling strategy will follow the line of fitting a base model in which we assume no difference between students but still to some degree adjust for differences between interviews by considering one or two interview based covariates. One can then try to expand the model structure by fitting a model in which we consider the whole nested structure of data possibly including a random effect on the level of students. This will be discussed briefly in section 5. The purpose of the base model is to establish an idea of which covariates to include in a more complex model. Another purpose is to see whether its even possible to fit a model at this simpler level due to the lack of non-correct outcomes in the response. But before we start modeling a short presentation of the model structure will be given in section 4.1 followed in section 4.1.2 by a formal definition of complete and quasi-complete separation. The reason for including the section dedicated to separation is because it is of importance when it comes to understand the problem of fitting a logistic regression model using maximum likelihood based inference when facing the type of data used in this thesis.

4.1 Logistic regression modeling

The response variable ”correct use” is assumed to be independent realizations of Bernoulli distributed variables, i.e $Y_{j,k,l} \sim B(\pi_{j,k,l})$. The subscripts $\{j,k,l\}$ are defined according to the nested structure as $j = 1,..,25$, $k = 1,..,n_j$, $l = 1,...,n_{j,k}$, where j refers to a specific student , k refers to a specific interview of student j and l refers to a specific noun adjective combination in this interview. To avoid this cumbersome but informative subscript notation we will from now on in this section use the single subscript i such that the response now can be written as $Y_i \sim B(\pi_i)$ where $i = 1,...,n$ and n is the total number of realizations. A standard approach for modeling binary data is the logistic regression model (Agresti, 2013), which is part of the family of generalized linear models. The logistic regression model structure can be written as

$$\text{logit}(\mathbb{E}(Y_i)) = \text{logit}(\pi_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}$$

where $\boldsymbol{\beta}^t = (\beta_0,...,\beta_p)$ is the vector of unknown parameters and $\boldsymbol{x}_i^t$ is the i'th row of the design matrix X of the model with dimension $n \times (p+1)$ where row $\boldsymbol{x}_i^t$ corresponds to the relevant covariates of observation y_i. Since the logit-link is just the log odds ratio, that is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

we get that the probability of correct use evaluated in the i'th realization is given by

$$\mathbb{E}(Y_i) = \pi_i = \frac{\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}. \tag{1}$$

4.1.1 Parameter inference based on the log likelihood

Consider the binary variable Y conditioned on the vector $\boldsymbol{x}^t = (x_0,...,x_p)$ where $\boldsymbol{x} \in \mathbb{R}^{p+1}$ represent an encoding of continuous and discrete covariates. Now, given the logit link from (1) we have that

$$P(Y=1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}^t\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}^t\boldsymbol{x})} = \frac{1}{1+\exp(-\boldsymbol{\beta}^t\boldsymbol{x})},$$

and hence

$$P(Y=0|\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{\beta}^t\boldsymbol{x})}.$$

Given our n observations we define the two sets $H_1 = \{i : Y_i = 1\}$, $H_0 = \{i : Y_i = 0\}$ where $i = 1, ..., n$. We may now write the log likelihood function as

$$\text{l}(\boldsymbol{\beta}; y, X) = \sum_{i \in H_1} \log\left(\frac{1}{1+\exp(-\boldsymbol{\beta}^t\boldsymbol{x}_i)}\right) + \sum_{i \in H_0} \log\left(\frac{1}{1+\exp(\boldsymbol{\beta}^t\boldsymbol{x}_i)}\right). \quad (2)$$

Using the method of maximum likelihood for inference of the parameter estimates $\hat{\boldsymbol{\beta}}$ we need to locate the point $\boldsymbol{\beta}$ which maximizes the log likelihood function. For logistic regression models in general, due the non linear relationship between the linear predictor $X\boldsymbol{\beta}$ and $\mathbb{E}(\boldsymbol{Y})$ this has to be done by implementing iterative numerical methods such as the IRLS algorithm (see Wood, 2006). By ”existence” of the maximum likelihood estimate we mean finitness and uniqueness. Under some circumstances due to structural properties of the design matrix of the model the estimate does not exist and hence the IRLS algorithm does not converge. We will now take a closer look at these properties and define the concept of complete and quasi complete separation. The reason for this detour is because in literature like Agresti (2013); Hosmer and Lemeshow (2000) the focus is mainly on complete separation caused by continuous covariates. The question of ”what kind of separation one may encounter if adding discrete covariates that considered one at a time does not generate any kind of separation?” is not addressed. The question is of importance when it comes to understand the problem we are facing when modeling the type of data considered in this thesis.

4.1.2 Complete and quasi-complete separation

The following definitions and theorems will just consider the case of separation in an arbitrary logit model with a binary outcome, which is a special case of the general definition given in Albert and Anderson (1984) and which we will follow in notation in this section. First we define a classification rule with which we can classify each row $\boldsymbol{x}_i$ of the design matrix X into two separate classes G_1 and G_0. We say that $\boldsymbol{x}_i$ is allocated in group G_1 iff

$$\boldsymbol{\beta}^t\boldsymbol{x} \geq 0 \quad (3)$$

and in group G_0 iff

$$\boldsymbol{\beta}^t\boldsymbol{x} \leq 0. \quad (4)$$

To clarify the above we can understand the classification rule as if $\text{logit}^{-1}(\boldsymbol{\beta}^t\boldsymbol{x}_i) \geq 0.5$ the row vector $\boldsymbol{x}_i$ will be allocated in G_1 and if

$\text{logit}^{-1}(\boldsymbol{\beta}^t\boldsymbol{x}_i) \leq 0.5$ the vector $\boldsymbol{x}_i$ will be allocated in G_0. In case of equality the vector will be allocated in both G_1 and G_0, that is $\boldsymbol{x}_i \in G_1 \cap G_0$. Now we can proceed with the definition.

Definition 1 *If there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ such that* $\forall i \in H_1, i = 1, ..., n$

$$\boldsymbol{\beta}^t\boldsymbol{x}_i > 0, \tag{5}$$

and $\forall i \in H_0$

$$\boldsymbol{\beta}^t\boldsymbol{x}_i < 0 \tag{6}$$

*there is **complete separation** of the sample points.*

As a consequence $H_1 = \{i : \boldsymbol{x}_i \in G_1\}$ and $H_0 = \{i : \boldsymbol{x}_i \in G_0\}$ and since $H_1 \cap H_0 = \emptyset$ it follows that $G_1 \cap G_0 = \emptyset$ and why we say there is complete separation. We can interpret the result as if the parameter vector $\boldsymbol{\beta}$ correctly allocates all $\boldsymbol{x}_i$ to the actual realizations of the Y_i. Given the existence of a vector $\boldsymbol{\beta}$ we see that the results holds for all vectors $k\boldsymbol{\beta}$ where $k \in \mathbb{R}_+$ and that $\forall i \in H_1$

$$0 < k\boldsymbol{\beta}^t\boldsymbol{x}_i \to \infty, \quad \text{when} \quad k \to \infty \tag{7}$$

and $\forall i \in H_0$

$$0 > k\boldsymbol{\beta}^t\boldsymbol{x}_i \to -\infty \quad \text{when} \quad k \to \infty. \tag{8}$$

Now if we evaluate (2) in $k\boldsymbol{\beta}$

$$\sum_{i \in H_1} \log\left(\frac{1}{1 + \exp(-k\boldsymbol{\beta}^t\boldsymbol{x}_i)}\right) + \sum_{i \in H_0} \log\left(\frac{1}{1 + \exp(k\boldsymbol{\beta}^t\boldsymbol{x}_i)}\right). \tag{9}$$

it directly follows from (7) and (8) that

$$\text{l}(k\boldsymbol{\beta}; y, X) \to 0 \quad \text{when} \quad k \to \infty. \tag{10}$$

That is given complete separation we get that 0 is the maximum of the log likelihood and that the estimate $\hat{\boldsymbol{\beta}}$ is given by a point on the infinite boundary of the parameter space and why we can conclude that no maximum likelihood estimate exist. This result we state as a theorem.

Theorem 1 *If there exists a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ which completely separates the set of data points, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ does not exist, and* $\max_{\beta \in \mathbb{R}^{p+1}} \text{l}(\boldsymbol{\beta}; y, X) = 0$.

An interesting interpretation of complete separation is that given the "non" existing estimate the model correctly predicts all outcomes of the response. This is related to the saturated model which given the definition is a special case of complete separation. Now let us define quasi-complete separation.

Definition 1 *If there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ such that* $\forall i \in H_1, i = 1, ..., n$

$$\boldsymbol{\beta}^t \boldsymbol{x}_i \geq 0, \tag{11}$$

and $\forall i \in H_0$

$$\boldsymbol{\beta}^t \boldsymbol{x}_i \leq 0 \tag{12}$$

with equality for at least one (i)*, there is* ***quasi-complete separation*** *of the sample points.*

The definition implies that G_1 and G_0 completely separates the sample points except for at least one i such that $\boldsymbol{x}_i \in G_1 \cap G_0$ why we say there is quasi-complete separation. For the quasi-complete case we have a similar result as the one for complete separation here stated without the proof (see Albert and Anderson, 1984). The proof is similar to the proof of theorem 1 and shows that the maximum is reached at the infinite boundary of the parameter space.

Theorem 2 *If there exists a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ which quasi-completely separates the set of data points, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ does not exist, and* $\max\limits_{\beta \in \mathbb{R}^{p+1}} \mathrm{l}(\boldsymbol{\beta}; y, X) < 0$.

When fitting the logistic regression model we will do so by calling the standard *glm* function while using R. As an extra step of precaution we will use the "safeBinaryRegression" package developed by Konis (2013) which before calling the *glm* function implements linear programming algorithms to test for separation.

4.1.3 Examples of separation

To get an idea of the consequences of the definitions when working with simple binary covariates let us take a look at a few simple examples. Assume a binary outcome $\boldsymbol{y}$ and the two binary covariates $\boldsymbol{x}_1, \boldsymbol{x}_2$.

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

To test for separation for x_1 we get to solve the linear inequalities given by $X\boldsymbol{\beta}$ and check whether there exist a $\boldsymbol{\beta}$ that separates the rows of X according to $\boldsymbol{y}$.

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \Longrightarrow \begin{array}{rcl} \beta_0 & > & 0 \\ \beta_1 & < & 0 \\ \beta_0 & < & -\beta_1 \end{array} .$$

The reason for keeping the vector $\boldsymbol{y}$ is to show how we are suppose to split the rows of the design matrix. We see that all vectors $\boldsymbol{\beta} \in \mathbb{R}^2$ satisfying the inequalities to the right completely separates the outcome. For $\boldsymbol{x_2}$ we get

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \Longrightarrow \begin{array}{rcl} \beta_0 + \beta_1 & \geq & 0 \\ \beta_0 & \leq & 0 \\ \beta_0 + \beta_1 & \leq & 0 \end{array} .$$

We see that the only solution here is all $\boldsymbol{\beta}$ such that $\beta_0 + \beta_1 = 0$ and $\beta_0 < 0$. These $\boldsymbol{\beta}$ generate quasi-complete separation of the sample points. These two examples are equivalent to all 2×2 tables of the form

$$\boldsymbol{x_1} \equiv \begin{vmatrix} a & 0 \\ 0 & c \end{vmatrix}, \boldsymbol{x_2} \equiv \begin{vmatrix} a & b \\ 0 & c \end{vmatrix}$$

where $(a, b, c) \in \mathbb{N}_+^3$. So if we in our univariate analysis encounters such tables they imply complete and quasi-complete separation, respectively. Now consider the following three covariates

$$\boldsymbol{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \boldsymbol{x}_4 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} . \boldsymbol{x}_5 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

of which none implies separation when used separately. First consider the model including $\boldsymbol{x_3}$ and then we add the covariate $\boldsymbol{x_4}$ which generates the following linear inequalities.

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \Longrightarrow \begin{array}{rcl} \beta_0 + \beta_2 & \geq & 0 \\ \beta_0 + \beta_1 & \geq & 0 \\ \beta_0 + \beta_1 + \beta_2 & \leq & 0 \\ \beta_0 + \beta_1 & \leq & 0 \\ \beta_0 & \leq & 0 \end{array} ,$$

which implies

$$\begin{array}{rcl} \beta_0 & = & 0 \\ \beta_2 & \geq & 0 \\ \beta_1 & \geq & 0 \\ \beta_1 + \beta_2 & \leq & 0 \\ \beta_1 & \leq & 0 \end{array} \implies \begin{array}{rcl} \beta_0 & = & 0 \\ \beta_1 & = & 0 \\ \beta_2 & \geq & 0 \\ \beta_2 & \leq & 0 \end{array} \implies \begin{array}{rcl} \beta_0 & = & 0 \\ \beta_1 & = & 0 \\ \beta_2 & = & 0 \end{array} .$$

The only solution is the null vector why we can conclude there is no separation. Now finally, let us put $\boldsymbol{x}_3, \boldsymbol{x}_5$ in the same model.

$$\boldsymbol{y} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \implies \begin{array}{rcl} \beta_0 & \geq & 0 \\ \beta_0 + \beta_1 + \beta_2 & \geq & 0 \\ \beta_0 + \beta_1 + \beta_2 & \leq & 0 \\ \beta_0 + \beta_1 & \leq & 0 \\ \beta_0 & \leq & 0 \end{array} ,$$

which implies

$$\begin{array}{rcl} \beta_0 & = & 0 \\ \beta_1 + \beta_2 & \geq & 0 \\ \beta_1 + \beta_2 & \leq & 0 \\ \beta_1 & \leq & 0 \end{array} \implies \begin{array}{rcl} \beta_0 & = & 0 \\ \beta_1 & = & -\beta_2 \\ \beta_1 & < & 0 \end{array} .$$

We see that all vectors $\boldsymbol{\beta} \in \mathbb{R}^3$ satisfying the inequalities to the right generates quasi-complete separation. That is, even though the two covariates does not imply separation when treated separately the combination of the two does. This kind of separation we cannot expect to discover using univariate analysis alone. As a further result, when in this case considering all possible binary covariates that doesn't imply separation on their own and then combining them with $\boldsymbol{x_3}$ four out of ten possible combinations implied quasi-complete separation and none complete separation. So when facing non convergence of the IRLS algorithm due to separation of the sample points and working with discrete covariates we are most likely facing quasi-complete separation.

4.2 Model-building

Now we are ready for some modeling. When adding new covariates we will use a forward selection strategy combined with the likelihood ratio test statistic. As proposed by Hosmer and Lemeshow (2000) a 15% level of significance for introduction and a 20% level of significance for excluding will be implemented. Which variable to include at each step is the one which generates the smallest p-value in terms of the likelihood ratio statistic.
Since the goal of this thesis is to answer the hypothesis in section 1.4 we

will when fitting the base model just consider the subset of variables that are of relevance for this purpose. The selected variables are summarized in the following table.

Level	**Abbreviation**	**Description**	**Type**	**Values**
Interview				
	vocd	VOCD	Continuous	$\mathbb{R}_+$
	months.rec	Time in months since first interview	Ordinal	$\mathbb{R}_+$
Noun				
Adjective	*numb*	Number	Binary	{sing, plur}
Combination	*gend*	Gender	Binary	{fem, masc}
	noun.reliab	Noun Reliability	Continuous	$[0, 1]$
	noun.avail	Noun Availability	Continuous	$[0, 1]$
	noun.valid	Noun Validity	Continuous	$[0, 1]$

4.2.1 Selection of variables

In the initial step of the model building process we conducted a stepwise forward selection. At each step we test for separation. The summary of the likelihood ratio square tests from each step are summarized in table 11 and the model specifications can be found in table 10. At each step no variable was up for exclusion and the only variable not included was *months.rec* generating a p-value > 0.15 corresponding to mod6 in table 11. The resulting multivariable model mod5 is summarized in table 12. Variable *numb* is modeled using plural as the level of reference and *gend* as using feminine as reference.

Model name	Predictor
mod0	1
mod1	1 + noun.reliab
mod2	1 + noun.reliab + vocd
mod3	1 + noun.reliab + vocd + numb
mod4	1 + noun.reliab + vocd + numb + gend
mod5	1 + noun.reliab + vocd + numb + gend + noun.avail
mod6	1 + noun.reliab + vocd + numb + gend + noun.avail + months.rec

Table 10: The seven investigated logit models and their linear predictor.

In table 13 we have the results of fitting a univariate logistic regression model to each of the included variables. Comparing these results with the multivariate results in table 12 we can see that *noun.avail* show signs of weaker association with the outcome, with a p-value $= 0,067$ of the Wald statistic, when combined with the other variables included. Looking at the estimate for *vocd* we see almost no change in both the estimate and the estimate standard error indicating independence.

So far there are three continuous variables *vocd*, *noun.reliab* and *noun.avail* in the model. In the univariate analysis none of these exhibited an apparent

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
mod0	3095	1879.60			
mod1	3094	1772.47	1	107.13	0.0000
mod2	3093	1745.17	1	27.30	0.0000
mod3	3092	1727.43	1	17.74	0.0000
mod4	3091	1722.22	1	5.21	0.0225
mod5	3090	1719.02	1	3.21	0.0734
mod6	3089	1717.21	1	1.81	0.1784

Table 11: The succescive likelihood ratio tests of the stepwise forward selection procedure.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6905	0.3172	-2.18	0.0295
noun.reliab	0.9222	0.4039	2.28	0.0224
vocd	0.0152	0.0030	5.12	0.0000
numbsg	0.9455	0.2388	3.96	0.0001
gendm	0.3632	0.1427	2.54	0.0109
noun.avail	0.7448	0.4060	1.83	0.0666

Table 12: Summary table for the log-odds ratio parameter estimates of mod5.

linear association with the rate of correct use. We now take a look at these variables on the logit scale to see whether linearity is justified and if not, if there exists some reasonable transformation that still admits understandable interpretation. As proposed by Hosmer and Lemeshow (2000) we start with the continuous variable with the lowest *p*-value of the Wald statistic and then work our way up the *p*-values. This means, *vocd* then *noun.reliab* and then *noun.avail*.

We start by plotting the LOESS curve of *corr.use* against *vocd* but now transformed to the logit scale. The plot in figure 11 does not look to promising. The curve seems to follow a sine wave and why there seem to be a problem of linearity in the logit scale. At the same time one should be aware of that the scale on the y-axis somewhat exaggerates the curvature. To what extent it is not linear is another question which needs to be answered using further methods which here follows. Let us investigate the two tails of the logit by using a design variable which we define by splitting the values of *vocd* in to relevant intervals of interest and consider these intervals as levels of a factorial variable. We then refit the model replacing *vocd* with its factorial correspondent. By using all covariates the interactions with the other variables in the model are taken into account. We then plot the different coefficients of the levels on the mid points of their intervals to get a picture of how the slope of the variable changes over the intervals. Why we do this

	Estimate	Std. Error	OR	2.5%	97.5%	G	p
noun.reliab	2.040	0.1969	7.69	5.24	11.34	107.13	0.000
vocd	0.016	0.0029	1.02	1.01	1.02	29.95	0.000
numbsg	1.030	0.1282	2.80	2.18	3.61	66.17	0.000
gendm	0.616	0.1300	1.85	1.44	2.40	23.30	0.000
noun.avail	0.696	0.2178	2.01	1.30	3.06	9.79	0.002
months.rec	0.029	0.0114	1.03	1.01	1.05	7.44	0.006

Table 13: Results of univariate logistic regression models for all variables where OR is the estimated odds ratio, a 95% highest likelihood confidence for the odds ratio, G the likelihood ratio statistic and p the corresponding p-value.

is to investigate whether the tails just might be the result of numerical properties of the LOESS curve. The results of the design variable are presented in table 14 and the corresponding plot in figure 12. The trend is similar to the plot in figure 11 with a decrease in both tails. The 95% confidence intervals in table 14 all include 0 why none of the factorial levels are significantly different from 0, which doesn't agree with the fact that *vocd* being significant in mod5. All confidence intervals for the parameters overlap and thereby indicating that the change with increasing values of VOCD is small relative to the scale.

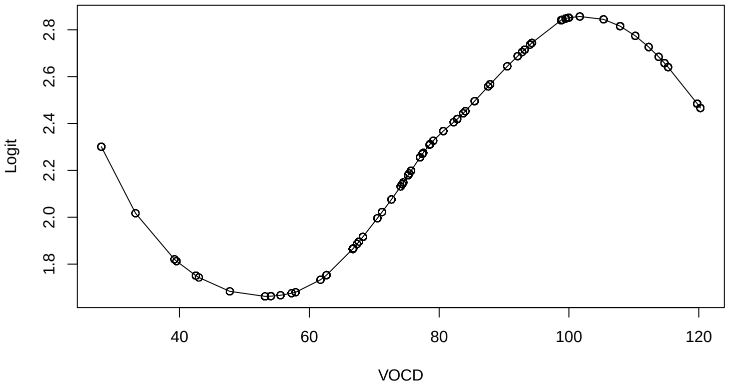

Figure 11: Transformed LOESS curve on logit scale between response and VOCD.

As a final step we use the method of fractional polynomials as presented in Hosmer and Lemeshow (2000). This method tests, given the set $\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ of powers, if there exists a transformation of the

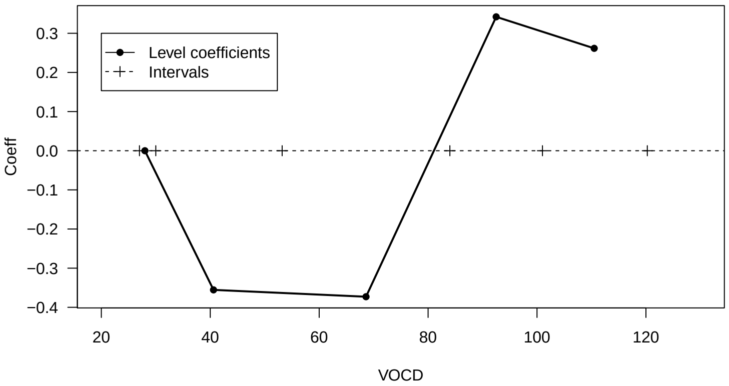

Figure 12: The coefficients of the log odds ratios from table 14 against the mid points of their intervals.

	Interval	Number	Coeff.	2.5%	97.5%
Group 1	0-28	43	0		
Group 2	28-53	232	-0.355	-1.64	0.68
Group 3	53-84	1232	-0.373	-1.61	0.58
Group 4	84-101	866	0.342	-0.91	1.33
Group 5	101-120	723	0.262	-0.99	1.25

Table 14: The design variable of the VOCD measure from the multivariable model mod5 presented in table 12.

continuous covariate given by a single or a dual combination of the powers $\wp$ that better fits the data under influence of the model. We test this by running trough all $J = 1$ single transforms given by $\wp$ and all $J = 2$ dual combinations of transforms given by $\wp$. We then compare the best $J = 2$ model, the one with the largest likelihood, with the original model and check if there is a significant improvement of the likelihood at the 5% level. If the test is not significant we keep the covariate as linear in the logit. Otherwise we compare the best $J = 2$ with the best $J = 1$ model and test if the $J = 2$ model is significantly better than the $J = 1$ model. If the model $J = 1$ is significantly better we use the $J = 2$ model with the given transform and if not we choose model $J = 1$. The method thereby tests for linearity on the logit scale and if there is proof of non linearity it suggests a transformation. What needs to be added is that by the power 0 we mean the $\log(\dot{})$ transform.

The results of applying the method on *vocd* is summarized in table 15. The reason for the value 0 of the likelihood ratio statistic between the best model $J = 1$ and the original model is because the best power chosen is 1 which is just the linear form identical to the original model. The p-value 0.059 for $J = 2$ is not significant at the 5% level why the fractional polynomial test approves of treating *vocd* as linear. Due to this last result we end up treating *vocd* as linear in the logit.

	df	Deviance	G	P-value	Powers
Linear	0	1719.016			1
J=1	1	1719.016	0	1	1
J=2	3	1711.558	7.457	0.059	-2, -1

Table 15: Summary of the fractional polynomials method for VOCD. G is the likelihood ratio statistic between the linear model and the fractional polynomial.

We now proceed with noun reliability using the same methods as we did with VOCD. In figure 13 we can see that there is a slight dip in the right tail. The p-values from the method of fractional polynomials in table 16 indicate that it's preferable to treat *noun.reliab* as linear in the logit why we choose to do so.

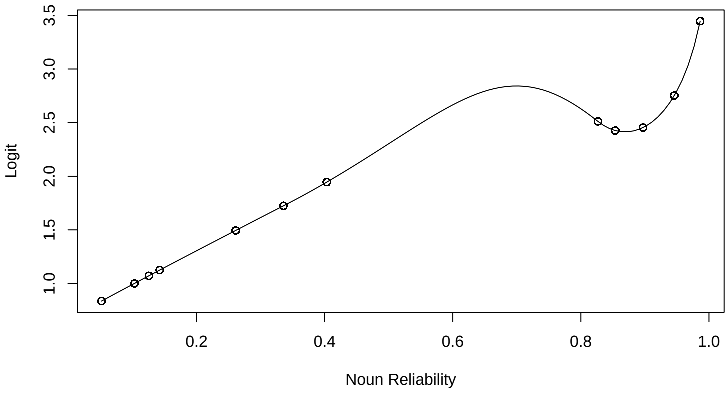

Figure 13: Transformed LOESS curve on logit scale between response and noun reliability.

	df	Deviance	G	P-value	Powers
Linear	0	1719.016			1
J=1	1	1714.992	4.024	0.045	log(.)
J=2	3	1714.353	4.663	0.198	-1,-1

Table 16: Summary of the fractional polynomials method for noun reliability. The log(.) in the power column represent log() transformation.

Finally we take a look at noun availability. As written in section 3 this covariate showed signs of strange behavior where the LOESS curve took values outside the interval $[0,1]$ (see figure 10). To be able to generate a plot in the logit scale we have to increase smoothness of the LOESS function which will generate a curve more linear in shape making the logit transform possible. The result we have in figure 14 where we can see that even though the LOESS curve has a higher degree of smoothness it still exhibits non linearity on the logit scale.

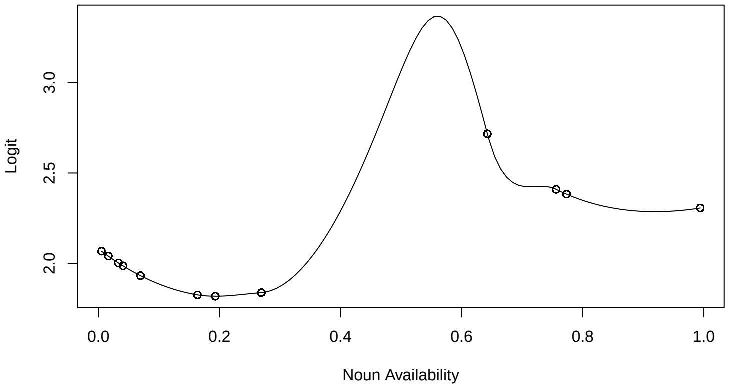

Figure 14: Transformed LOESS curve on logit scale between response and noun availability.

Now we examine the linearity using the method of design variables. We split *noun.avail* according to the quartiles but due to the uneven distribution of *noun.avail* we choose to split the interval of the first quartile in two. We see in figure 15 and in table 17 the corresponding estimates of the coefficients of the log odds ratio of each interval. As expected the plot do not show signs of linearity and given the 95% confidence intervals in table 17 there is strong evidence of non linearity since none of them contains 0. A possible transformation suggested by these results is treating noun availability as a categorical variable dividing it according to the intervals in table 17 but merging the last two intervals together since the estimates of these are very similar and the corresponding interval greatly overlap.

The result of the likelihood ratio statistic between the model mod7 in which we use a 4-level categorical version of noun availability and mod4, the model without noun availability, we have in table 18. Comparing the resulting p-value close to 0 with the original p-value of 0.073 given in table 11 we see that we get a model which is significantly better than the model in which we keep noun availability as linear in the logit.

As a final step we use the method of fractional polynomials. The results

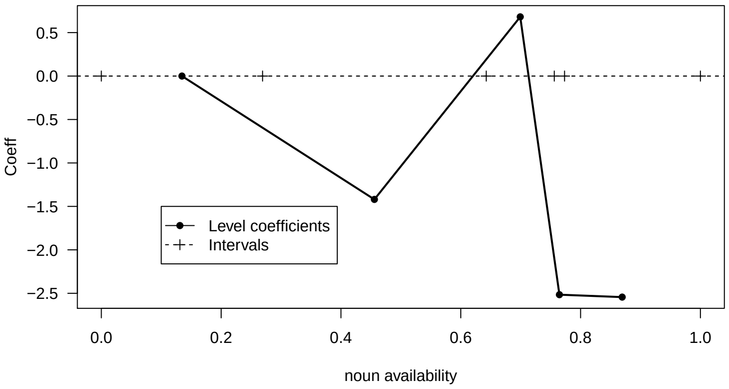

Figure 15: The coefficients of the log odds ratios from table 17 against the mid points of their intervals.

	Interval	Number	Coeff.	2.5%	97.5%
Group 1	0-0.26	558	0		
Group 2	0.26-0.64	774	-1.42	-3	-0.12
Group 3	0.64-0.75	478	0.681	0.01	1.32
Group 4	0.75-0.77	711	-2.516	-4.67	-0.73
Group 5	0.77-1	575	-2.544	-4.93	-0.53

Table 17: The estimated effects of the categorical version of noun availability in model mod5 presented in table 12.

in table 19 propose that we treat noun availability as non linear but comparing the deviance of the model with J= 2 and the deviance of the model mod7 in which we treat noun availability as categorical we see that the later mod7 is the better choice.

To sum up we ended up by treating noun availability as categorical with four levels and we will define it as the variable *avail*. In table 20 we have the resulting model which we labeled mod7. We can see that the estimate of the coefficient of number no longer is significant and both gender and noun reliability are now highly significant. Treating noun availability as categorical thereby had some major effects.

Looking at the definition of noun availability in section 2.1.4 we can see that we could expect some interaction between noun availability and gender and noun availability number since this measure is defined in terms of gender and number. Controlling for interactions between noun availability and gender and number respectively, we in both cases get separation. That is, no maximum likelihood estimate exists of the parameters given that we expand

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
mod4	3091	1722.22			
mod7	3088	1692.59	3	29.63	0.0000

Table 18: The likelihood squre test of comparing the model where we use the categorical rescaling of availability to mod4.

	df	Deviance	G	P-value	Powers
Linear	0	1719.016			1
J=1	1	1718.146	0.87	0.351	-2
J=2	3	1699.627	19.389	0	0.5, 0.5

Table 19: Summary of the fractional polynomials method for noun availability.

the model with one of the interactions. We could say that we somewhat reached a dead end using the method of maximum likelihood estimation and thereby we consider mod7 as our final base model. The discussion contains ideas concerning the further analysis. Lets check the fit of mod7 before we interpret some of its model parameters.

4.2.2 Model checking of the final model

When dealing with binary outcomes the deviance between the saturated model and a given model of the binary outcome can not be considered asymptotically chi-square distributed since the number of parameters in the saturated model are not fixed. We here use instead the Hosmer-Lemesow goodness of fit test dividing the fitted values into 10 batches since we have a total of 8 parameters in mod7.

The test generates a p-value of 0.241 indicating no lack of fit at a 5% level of significance. The binary response makes ordinary model checking quite difficult but here we use a parametric bootstrap approach creating an R function by using and improving code written by Wood (2006). The function is used in two ways. First it creates a 95% bootstrap envelope for the cumulative distribution of the residuals under the hypothesis that the given model is correct. Secondly it uses the same bootstrap samples to calculate the number of runs of each sample, that is the number of runs of independent observations given the fitted model. Using the runs of each sample and the number of runs of the observed data it generates an approximate p-value under the hypothesis of independence. In figure 16 to the left the empirical CDF with the 95% bootstrap interval indicating that our distributional assumptions are sensible. To the right we have the simulated runs of a total of 200 samples confirming the p-value of 0 under the null hypothesis of independence in the residuals. This is not surprising since we are under

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8449	0.4050	-4.56	0.0000
avail2	-1.4101	0.7215	-1.95	0.0507
avail3	0.6940	0.2824	2.46	0.0140
avail4	-2.5079	0.9966	-2.52	0.0118
noun.reliab	4.5999	1.2174	3.78	0.0002
numbsg	0.4635	0.2851	1.63	0.1040
gendm	1.4748	0.3367	4.38	0.0000
vocd	0.0152	0.0030	5.05	0.0000

Table 20: Coefficient summary table for the log odds ratios in mod7.

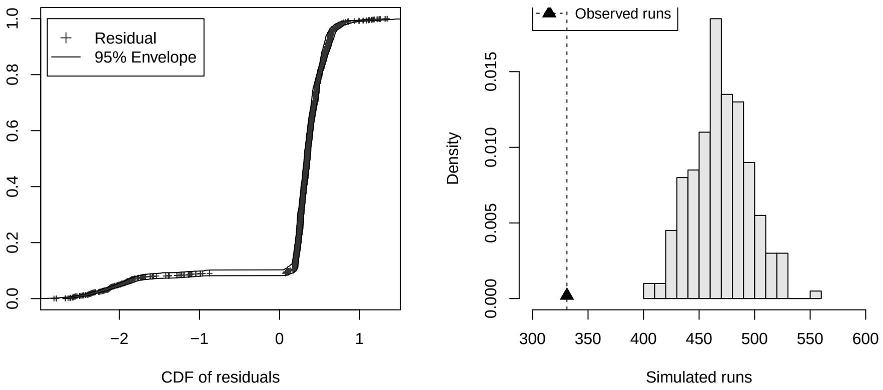

Figure 16: Left: The empirical CDF of residual with 95% bootstrap envelope. Right: Histogram of simulated runs with the observed runs represented as a point.

the base model assuming no difference between students which of course is rather naive. The high dependence of the residuals suggests that we need to consider those differences that is considering the whole nested structure. Another possible cause for the lack of independence is the absence of relevant variables we do not have access to.

4.2.3 Parameter interpretation of the final model

In this section we will give a short interpretation and presentation of the results of the final model mod7 and the way they relate to the hypothesis. A summary of the estimates of the final model mod7 including the odds ratios and their respective 95% highest likelihood confidence intervals can be found in table 21.

	Estimate	Std.Err	z	P>|z|	OR	2.5%	97.5%
(Intercept)	-1.84	0.40	-4.56	0.00	0.16	0.07	0.34
avail2	-1.41	0.72	-1.95	0.05	0.24	0.05	0.88
avail3	0.69	0.28	2.46	0.01	2.00	1.13	3.44
avail4	-2.51	1.00	-2.52	0.01	0.08	0.01	0.48
noun.reliab	4.60	1.22	3.78	0.00	99.48	11.37	1331.78
numbsg	0.46	0.29	1.63	0.10	1.59	0.90	2.75
gendm	1.47	0.34	4.38	0.00	4.37	2.31	8.66
vocd	0.02	0.00	5.05	0.00	1.02	1.01	1.02

Table 21: The estimated coefficients including the estimates of the odds ratios (OR) and thier corresponding 95% highest likelihood intervals.

We first consider the hypothesis that the rate of correct use will increase with higher values of VOCD. Due to size of the estimated OR we will give the estimate on the scale where one unit represent 1 standard deviation of VOCD. The odds ratio is telling us that the odds of correct use increase with a factor of 1.38 for each standard deviation increase in VOCD. Suggested by the confidence intervals this change can be small as 1.22 or large as 1.57 with a 95% confidence. The change is overall small but still confirms the hypothesis of an increase in the rate of correct use given an increase in VOCD.

Now looking at the hypothesis that higher values of noun reliability and noun availability will have a positive effect on the outcome we can tell from the estimates of the levels of *avail* that the model does not support this assumption in the case of noun availability. Under the influence of the other covariates it actually have a negative effect why we under mod7 need to reject the hypothesis that noun availability has a positive effect. Now if we consider noun reliability we have an odds-ratio estimate of 99.48. This ratio is quite hard to interpret since noun reliability is bounded on the closed interval $[0, 1]$ why we choose to give the odds ratio for a 0.1 point one increase in noun reliability. For each increase in noun reliability of 0.1 the odds of correct use multiply with a factor of 1.58 with a 95% confidence interval $(1.28, 2.05)$. This confirms the hypothesis that higher values of noun reliability have a positive effect on the outcome.

That masculine agreement means higher rates of correct use we can see by looking at the corresponding estimate of the odds-ratio. When considering masculine gender the odds multiply by a factor of 4.37. The 95% highest likelihood intervals indicate that this change can be as small as 2.307 or as large as 8.665. This confirm the hypothesis that masculine gender will have higher rates of correct use. Looking at number agreement we have a point estimate 1.59 of the odds ratio when considering singular agreement. The confidence interval $(0.898, 2.753)$ contains the value 1 why we cannot

confirm the hypothesis of a strictly positive effect.
Finally we look at the last hypothesis that learners of Italian as a second language get better on using number and gender agreement over time. This hypothesis we need to reject since the variable *months.rec* was not included in the final model. The reasons why *months.rec* did not end up in model the will be addressed in the discussion.

5 Discussion

The aim of this thesis was to answer the hypothesis given in section 1.4 by a statistical analysis. Due to the sparse number of incorrect outcomes of the response we faced the numerical issue of separation when fitting the relevant logistic regression model. This in combination with the nested structure of data led to the strategy of first trying to fit a base model considering the covariates at the level of noun gender combinations and some covariates at the level of interview. The purpose of this strategy was to get an idea of which covariates to include in a more complex model considering the whole nested structure. Another purpose was to see whether it was even possible to fit a simple base model. As a result we ended up with a final base model which we named mod7 presented in table 21. When trying to extend the model by introducing interactions between the categorical version *avail* of noun availability and number and gender we immediately faced separation. We thereby were not able to proceed with a more extended analysis at least not when using maximum likelihood inference. In a analysis taking the whole nested structure into account it appears advantageous to use Bayesian inference. Two reasons for this are that if necessary we can add a minimal amount of information within the prior distributions of the parameters and that it is in some sense more natural to build hierarchical models using Bayesian inference (Carlin and Louis, 2009). Another reason for a bayesian approach is that when we have defined the model structure the inference can then be carried out by using statistical softwares such as Jags (Plummer, 2003).
When fitting the hierarchical model it might be a good a idea to consider variations in the students by including this into the model using random effects. Hence the student effects are nuisance parameters This is motivated by the fact that we are not primarily interested in specific differences between students but rather interested in some general linguistic characteristics concerning number and gender agreement. Another random effect of interest is one that tries to model the way students are included in relation to there learning phase. We could see in figure 3 that a majority of students showed no signs of improvement. This is most likely caused by the fact that most student already reached a significant level of Italian speech and thereby show no signs of improvement. This might explain why *months.rec* did not end up in the base model.

In the base model we ended up with a categorical re-scaling of the the continuous variable noun availability. There are good reasons to be skeptic of noun availability being a proper continuous predictor due to the way it is defined. Looking again at the definition in section 2.1.4 we can see that the measure defines how common a specific noun ending is conditioned on a specific number gender combination. That is, for the measure to have some general application the different groupings in terms of the number and gender combinations need to be independent. From its distributional characteristics we can conclude that this is not case. The noun availability measure thereby just makes sense by considering it within the respective groups. Intuitively the measure is a good idea and its behavior tells us that the different groupings are of importance but the results are hard to interpret with respect to the intuition behind the measure. Noun reliability has a similar problem but exhibits signs of independence. The only problem was the u-shaped part in the right tail. The cause of the u-shape is due to the fact that in the right tail we have a large number of different number and gender combinations that are unique within their specific noun-ending group and they thereby get a high reliability even though they are quite uncommon. In further studies a good idea would be be to define more general frequency measures not conditioned on different groupings. Questions about different groupings can then, within a model context, be examined by introducing the relevant interactions.

While fitting the model we encountered two major problems in this thesis: First the problem of separation to which we dedicated a number pages in order to get a better understanding of the actual problem. At each step of the model building we then checked for separation using the the ”safeBinaryRegression” package developed by Konis (2013) in R. The second problem was checking the model fit and residual diagnostics when dealing with binary data. For the model fit we used the Hosmer-Lemeshov test statistic that showed no signs of a overall lack of fit. To deal with the residuals we implemented a parametric bootstrap creating an R function by using and improving code given in Wood (2006). The distribution of the residuals were well contained within 95% envelope thereby confirming our distributional assumptions but the test for independence showed great lack of independence. Thereby our conclusions given by the model interpretations in section 4.2.3 should be cautious.

To sum up we can see that there are several reasons why to consider consulting a statistician when dealing with these kind of linguistic studies. First of all the structure and distributional characteristics of data are of a complex nature which cannot be ignored if one aims for good model fit. Separation and hierarchical modeling with random effects are modeling issues that presuppose a great extent of mathematical knowledge that can be provided by the statistician. Secondly, as was mentioned above, the statistician can be helpful when defining variables such as the frequency measures to make sure they are interpretable within model context. The linguist and statistician can then together interpret and discuss the results of the statistical analysis.

References

Agresti, A. (2013). *Categorical data analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Bardel, C. (2004). Il progetto interita: l´apprendimento dell´italiano l2 in un contesto svedese. In Erman, B., Falk, J., Magnusson, G., and Nilsson, B., editors, *Second Language Acquisition*, number 13 in Stockholm Studies in Modern Philology New Series, pages 11–30. Almqvist & Wiksell International, Stockholm.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian methods for data analysis.* Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74(368):829–836.

De Mauro, T., Mancini, F., Vedovelli, M., and Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato.* Etatslibri, Milano.

Gudmundson, A. (2012). *L´accordo nell´italiano parlato de apprendanti universitari svedesi: Uno studio sull´acquisizione del numero e del genere in una prospettiva funzionalista.* PhD thesis, Stockholm University.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression (Wiley Series in probability and statistics).* Wiley Interscience, New York; Chichester, 2 edition.

Konis, K. (2013). *safeBinaryRegression: Safe Binary Regression.* R package version 0.1-3.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk.* Lawrence Erlbaum Associates, Mahwah, NJ.

MacWhinney, B. (2005). A unified model of language acquisition. In Kroll, J. F. and de Groot, A. M. B., editors, *Handbook of bilingualism:psycholinguistic approaches*, pages 49–67. Oxford University Press, New York.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.*

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Wood, S. N. (2006). *Generalized additive models.* Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL. An introduction with R.

Xie, Y., Andrew, A., Zvoleff, A., Diggs, B., Pereira, C., Wickham, H., Jeon, H., Arnold, J., Stephens, J., Hester, J., Cheng, J., Keane, J., Allaire, J., Toloe, J., Takahashi, K., Kuhlmann, M., Caballero, N., Salkowski, N., Ross, N., Vaidyanathan, R., Cotton, R., Francois, R., Brouwer, S., de Bernard, S., Wei, T., Lamadon, T., Torsney-Weir, T., Davis, T., Zhu, W., and Wush (2013). *knitr: A general-purpose package for dynamic report generation in R.* R package version 1.5.