CHAPTER SEVEN

SOCIOCULTURAL DETERMINANTS OF GRAMMATICAL TABOOS IN GERMAN*

RALF VOGEL
UNIVERSITY OF BIELEFELD, GERMANY

1. Introduction

This paper continues work that has been presented in (Vogel to appear). It covers empirical research on a class of morphosyntactic phenomena that I called *grammatical taboos*. These are morphosyntactic constructions which are subject to *stigmatisation*. They typically occur in standard languages like standard German (SG), the language explored in this paper. Linguistic theory mainly ignores such phenomena as artificial distortions of "natural" languages. In experimental morphosyntax, they are usually seen as potential confounds which have to be avoided in the creation of stimulus material. In the research presented here and in previous work to be introduced below, I am trying to break with this tradition and subject grammatical taboos to systematic experimental testing.

Section 2 introduces the theoretical background and the main results of the study presented in (Vogel to appear). This current paper is devoted to the sociocultural dimension. Section 3 summarizes results from sociolinguistic research on relevant socioeconomic and sociocultural variables in standard language and presents the results from a pilot study

* The research and analyses presented in this paper have grown over several years. Parts of this work have been presented in different versions at different stages of the development of this paper at the universities of Bielefeld and Leipzig, the university of Wuppertal and the Minsk State Linguistic University, Belarus. I want to thank the audiences of these presentations for very helpful comments and suggestions. My special thanks goes to my student assistants, Ann-Christin Broschinski, Sina Münstermann and Judith Sieker, without whom empirical studies like the one presented here could not be undertaken.

on the effects of stereotypes about the use of stigmatised language among university students.

2. Grammatical taboos: Concept and previous results

In the German speaking communities (including a huge part of the linguistics community), the German standard language (*Standard German*, SG) is perceived as the prestige variant used primarily by persons with higher education, and/or for more prestiguous purposes, for instance in formal written communication.

In contrast to this prevailing ideological disposition, Standard German is nowadays used dominantly in everyday oral conversation among all social groups – but in an informal register that has been developing over the last fifty years. As it is a rather recent development (and perhaps due to elitist reservations against ordinary language), descriptive grammars of SG up to now do not include a systematic description of this informal register.

The informal register of SG diverges from its formal register in a number of details.[1] Standard German (historically *New High German*, NHG) has been for about 400 years a variety only used in written language, mostly formal written language. Its development was carried forward by small elite of literates, actually a small minority of the German speaking community.[2] Until the middle of the 20th century, oral communication mainly took place in regional dialects – a situation that is still to be found in the German speaking parts of Switzerland.

As Weiß (2005, 13) and others describe in detail, this special situation allowed for the introduction of all kinds of artefacts into the language system of Standard German which prevail to the present days. What also prevails, as discussed in (Vogel submitted), is a reservation towards spoken language that I identified as a core principle of the German standard language ideology, the general standard language taboo:

[1] The amount of these divergences also seems to increase because of the higher dynamics in the development of spoken varieties as compared to the usually more conservative written varieties. Whereas the pre-WWII situation was roughly a situation of diglossia, in the second half of the 20th century spoken and written language were very close, as informal language was based on the written standard. The 21st century, however, might envisage a renewed and growing tension between formal and informal language which is now a tension between registers within the same language, rather than the earlier diglossic tension between dialect and standard.

[2] The minority of literate people made up less than 5% of the population in the beginning of the 16th century (von Polenz 2000, 128). Of course, there also was an urban-rural discrepancy, with literates up to 10% in towns (ibid.).

(1) **General standard language taboo (GSLT)**
Don't write like you talk!

The GSLT had an important sociocultural function in establishing a supraregionally comprehensible written variety, but today it has lost its motivation due to the decline of the traditional regional dialects. Today's reality is that formal and informal language *overlap* by, say, 98%.

Given this huge overlap, adherence to the GSLT today seems impossible. In practice, it boils down to awareness to a not so large set of *shibboleths*: with respect to *grammar*, certain aspects of language are selected as being reserved for only speaking or writing. These aspects enjoy high attention by the speech community. I label those shibboleths as *grammatical taboos* and *grammatical zombies*, respectively:

(2) **Grammatical taboo (GT)**
A certain grammatical aspect of informal oral language must not be used in formal written language.

Grammatical zombie
A certain grammatical aspect of the (inherited) written language must be used, although it might not or no longer be part of informal spoken language and perhaps it even contradicts the grammatical principles of the current standard language.

Speakers usually do not treat different registers of a language equally, but privilege the more prestigious formal written register. Use of grammatical taboos in spoken language, if it is considered to be allowed at all, is seen by those speakers as usage of incorrect language. This leads to a paradox within the grammatical system of the language, as described in (3).

(3) **Paradox of grammatical taboos**

1. A taboo in a language L can only hold over a construction C, if C *exists*. Thus, C must be part of L's language system. Even more so, the general principles of L are such that C follows consistently from them.
2. Because of the taboo over C, speakers of L who conform to the taboo nevertheless *believe* that C does not belong to L.

Grammatical taboos can be differently salient. In (Vogel to appear), I compared two presumably stronger taboos with two presumably weaker ones. These are exemplified in (4) and (5):

(4) Salient grammatical taboos (explored in the experiment)
a. Maria tat ein Buch lesen. auxiliary *tun*
M. did a book read
b. Die Straße ist nass, weil es hat geregnet. *weil* V2-clause
the street is wet because it has rained

(5) Non-salient grammatical taboos (explored in the experiment)
a. Als Peter kam, hat Max bereits double perfect
when P. came has M. already
gewonnen gehabt
won had
b. Als Paul kam, neckte ich den *d*-pronoun
when P. came teased I him

Linguistics is interested in the native variants of speakers which they have acquired early in their life. We are not interested in linguistic ideologies and the extent to which speakers follow them. However, the usual participant of a linguistic elicitation study is an adult native speaker who has been confronted with the GSLT and its effects at school and in public for many years. It would be naïve to expect their acceptability judgements not to be influenced by this experience.

Grammatical taboos, which are reserved for spoken language at most, have a negative connotation due to the prescriptive discourse on language at school and in public, where the idea of linguistic correctness is highly connected with the prestige variety. This negative connotation can be measured as reduced acceptability in elicitation experiments.

Of course, this is the reason why we usually avoid anything that smells like stigmatisation in our stimulus material for elicitation experiments. However, this also means that precisely these kinds of phenomena are not being explored empirically. That way, stigmatisation is taken over silently by the linguists, as a kind of self-fulfilling prophecy.

One focus of the studies that I am carrying out is the question, to what extent stigmatisation can be neutralised in morphosyntactic elicitation experiments. In (Vogel to appear), I tested three different judgement types. An aesthetic and a norm-oriented judgement type (i.e., "is this beautiful/good language?", types "A" and "N", respectively), which are both close to a prescriptive attitude towards language, are contrasted with a "possibility" judgement ("is this possible in German?", type "P") that comes closest to what linguists are interested in.

A second methodic point is connected with the question how to decide the grammaticality status of grammatical taboos. For this purpose, I introduced the concept of *empirical grammaticality* as defined in (6):

(6) **Empirical Grammaticality**
The empirical grammaticality of some expression E_i in a language L is the *probability* $p(E_i,L)$ of E_i being judged as a *possible expression of L* by a speaker of L.

Empirical grammaticality is the subject matter of morphosyntactic elicitation experiments the results of which are used to *estimate* the empirical grammaticality of a morphosyntactic construction. Grammatical theories likewise are *models* of empirical grammaticality, from which hypotheses can be derived for acceptability studies. Researchers not only distinguish grammatical and ungrammatical, but treat grammaticality as *gradient*, distinguishing unmarked from light or more severe markedness. This further partitions the class of grammatical sentences. In (Vogel submitted), I have been able to confirm by and large the following hypothesis about *absolute acceptability*:

(7) **General hypothesis about *absolute* acceptability**

Sentences are expected to be judged as acceptable according to their degree of grammaticality, as given in the following table:

% acceptable	**Category**
90 − 100 %	✓ — unmarked
60 − 80 %	? — slightly marked
20 − 50 %	?? — marked
0 − 10 %	* — ungrammatical

In the study presented in (Vogel to appear), the four investigated grammatical taboo phenomena each are rated as marked phenomena. In addition to this hypothesis about absolute acceptability, I have been able to confirm a hypothesis about *relative acceptability*. It is related to the concept of *effect size* from inferential statistics which will also be used below.

A simple measure of effect size could be the difference in acceptability between two constructions. In inferential statistics, *standardised* effect size measures are mostly used, for instance *Cohen's d* (Cohen 1988), which is the difference between two means, divided by their pooled standard deviation. A *d* of 0.5 or higher counts as at least a *medium size*

effect which in Cohen's words is "likely to be visible to the naked eye of a careful observer" (Cohen 1992, 156). As linguists' estimations of the four levels of (un-)grammaticality are quite reliable, even without empirical investigation, contrasts between those categories might have at least medium effect size. In (Vogel submitted), this is formulated as *general hypothesis about relative acceptability*:

(8) **General hypothesis about *relative* acceptability**
Sentence types with different grammaticality status (✓,?,??,*) contrast at least with medium effect size in the direction "✓" > "?" > "??" > "*".

The study in (Vogel to appear) is a written questionnaire study where sentences are judged along a 7-point rating scale. The experiment is divided into three sub-experiments using an aesthetic, a normative and a possibility judgement, respectively. For the estimation of absolute acceptability, the rating scale is projected onto the probability scale with 0 and 1 as minimum and maximum. Table 1 displays the mean ratings for each of the four taboo phenomena under each judgement type.

	A	N	P	P − A
aux. *tun*	0.137	0.165	0.254	0.117
weil V2-clause	0.146	0.260	0.375	0.229
d-pronoun	0.219	.294	0.398	0.179
double perfect	0.231	0.258	0.403	0.172

Table 1: Means for the four grammatical taboo phenomena under each judgement type, plus the difference between the means for judgement types P and A for each taboo phenomenon (Vogel submitted)

The two scale effects of judgement type (beautiful language < norm-oriented language < informal language) and salience of the taboo are confirmed by the fact that from each cell, the cell rightwards and downwards in Table 1 increases – with the exception of the gray shaded cell, but this exception only concerns the contrast between the two non-salient taboos about which no hypothesis is postulated.

Another crucial finding is that grammatical taboo phenomena differ from ordinary grammatical markedness in the patterns of between-subject

variance. This difference is illustrated in Figure 1. For this figure, the mean ratings for each subject (from the subexperiment with judgement type P) for grammatical taboos (combined) and ordinary syntactic markedness have been calculated and mapped onto the underlying 7-point rating scale.

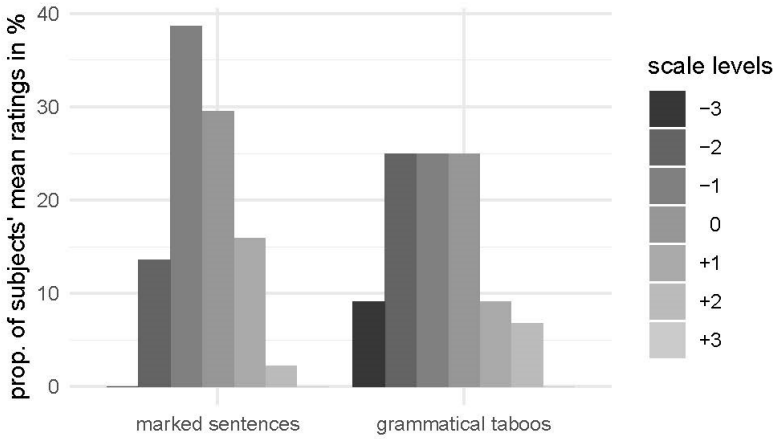

Figure 1: Proportional distribution of subjects' mean ratings (N = 44) over the seven scale levels for marked sentences and grammatical taboos, judgement type P

We see that the distribution of subjects' mean ratings for grammatical taboos is flatter and wider. The difference is too large to assume that the populations from which the samples are taken are nevertheless homogeneous in their variance. This is an important condition for the use of many tools of inferential statistics, including the effect size measure Cohen's d, introduced above. Therefore, a different measure of effect size is used in (Vogel submitted), *Cliff's delta* (Cliff 1996).[3] According to the

[3] Cliff's delta is the *difference* of the probabilities a) that an outcome from the first data set has a higher value and b) that an outcome from the second data set has a higher value. For its calculation with independent data sets all pairs (x_i, x_k) are built pairing each value from the first with each from the second data set. For each of these pairs it is determined whether x_i is higher or x_k, or whether they are equal. The total numbers of these three outcomes are counted and Cliff's *delta* is then calculated as the difference between the totals of pairs where one or the other has a higher value, divided by the total number of pairs. Under the convention that in calculating the numerator of this ratio the smaller value is always subtracted from

widely accepted interpretation of Cliff's delta by Romano et al. (2006), $d > .147$ is a small effect size, $d > .333$ a medium effect size and $d > .474$ is a large effect size.

Using Cliff's delta as measure of effect size and t-tests with Welch approximation for significance tests of data with different variances, it could be established that each of the four taboo phenomena contrasts with large effect size from ordinary ungrammatical sentences under the judgement type P – and this contrast is highly significant.[4]

This again confirmed our hypothesis on relative acceptability. Furthermore, it could be shown that this contrast breaks down under judgement type A for salient taboos, showing that they have a particular aesthetic disadvantage. This is summarised in Table 2.

	A	N	P
auxiliary *tun*	0.133	0.366 *	0.477 ***
weil V2-clause	0.089	0.488 ***	0.714 ***
d-pronoun	0.581 ***	0.513 ***	0.659 ***
double perfect	0.558 ***	0.718 ***	0.674 ***

Table 2: Pairwise comparisons of grammatical taboo violations and ungrammatical sentences for all taboo phenomena and judgement types: paired Cliff's delta and (Bonferroni-corrected) indicators of significance level for Welch's t-tests.
Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

The grey shaded cells in Table 2 signal medium to large effect sizes which are large enough to indicate a categorical difference between a taboo phenomenon and an ungrammatical, but semantically and syntactically equivalent sentence without taboo violation.

Overall, I could show in (Vogel to appear) that grammatical taboos, weak and strong ones alike, have a marked status empirically. This is due to stigmatisation based on the GSLT. Strong taboos have a particular aesthetic disadvantage, which does not show up with weak taboos.

the larger one, Cliff's delta takes on a value between 0 (no difference in probabilities) and 1 (maximum difference in probabilities).

[4] In each case, minimal pairs of a grammatical taboo phenomenon in eight lexical variants and a semantically and syntactically equivalent sentence without taboo violation, but with a verbal agreement error, have been compaired.

Auxiliary *tun*, the strongest taboo, differs from the other three phenomena in that it has much lower ratings across the board.

The participants of the experiment have been selected from a very homogeneous group: first year students of German studies at the University of Bielefeld. About three quarters of them were female, as is typical of study programmes in the humanities. Most of the students of the Bielefeld University come from its region and have made their high school degrees at one of a small number of secondary schools in the region. The high between-subject variance in judging taboo phenomena that we find here suggests that there are hidden sources of variation. The question, whether these could be of a socioeconomic or sociocultural origin, is the subject of this paper.

In section 3, insights from sociolinguistic research on the use of standard language by different social groups are discussed. The section also introduces the results of a pilot study on sociocultural stereotypes associated with the use of taboo phenomena by first year students of German studies.

Section 4 presents an acceptability judgement study that is designed to replicate the findings in (Vogel to appear), continues the discussion on the methodology of morphosyntactic experimentation by using a different experimental setting, and most importantly studies the impact of a number of sociocultural factors on the acceptability of grammatical taboos.

3. Gender and class-specific stereotypes about the use of grammatical taboos

3.1 The Gender Paradox

Several decades of sociolinguistic research established a robust finding about differences between males and females with respect to their use of (non-)standard language: “In virtually all sociolinguistic studies that include a sample of males and females, there is evidence for this conclusion about their linguistic behaviour: women use fewer stigmatized and non-standard variants than do men of the same social group in the same circumstances.” (Chambers 2009, 116)

It is important to avoid a false impression that such a condensed description might trigger: first, the difference between males and females is usually not large; second, it is only found *within* the same social status groups and in comparable contexts.

The effect is largest, in the words of Labov, in the “second highest group”: “As social awareness of a given change in progress develops, the

tendency of women to conform to conservative norms is exaggerated for the second highest status group (in North America and Western Europe, the lower middle class; in occupational terms, clerks, primary school teachers, small shopkeepers)." (Labov 2006, 320)

Labov (1990, 2006) couples these observations with the dimension of linguistic change. Some central insights of his own research spanning several decades of work and the work of many others are formulated in terms of four Principles in (Labov 2006):

> **Principle 1**: Linguistic change from below originates in a central social group, located in the interior of the socioeconomic hierarchy. (Labov 2006, 188, *the curvilinear principle*)

> **Principle 2**: For stable linguistic variables, women show a lower rate of stigmatized variants and a higher rate of prestige variants than men. (Labov 2006, 266, *the linguistic conformity of women*)

> **Principle 3**: In linguistic change from above, women adopt prestige forms at a higher rate than men.
> (Labov 2006, 274, "another aspect of the hypercorrect behaviour of the second highest status group")

> **Principle 4**: In linguistic change from below, women use higher frequencies of innovative forms than men do.
> (Labov 2006, 292)

The principles 2, 3 and 4 lead to what (Labov 2006) calls a *gender paradox*:

> **Gender paradox** Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not. (Labov 2006, 293)

I have a small reservation against the formulation of the paradox that Labov gives. It implies a particular interpretation of Principle 4 and the findings leading to it. When we focus on linguistic change we limit our view on those innovative forms that we have been able to identify as *successful* in linguistic change, i.e. new forms that have become part of the standard or at least are being used regularly to some degree already.

It is actually those successful innovative forms that have been investigated and for which it can be stated that women use them more frequently. But it could very well be that men are as innovative as women. The difference would then lie in the sad fact that their innovations are

picked up by others more rarely than women's innovations. In fact, Principle 2 suggests that men care less about the standard than women, so it would be surprising if their use of non-standard forms were restricted to stigmatised variants.[5]

How, after all, do we get to know that a particular newly observed variant is an instance of language change? The most important criterion is frequency of use, and, if it is possible for us to observe it, rising frequency of use. Unless a variant is observed with a particular frequency, we cannot even identify it as an instance of language change. We are literally blind for it.

So, I have my doubts whether women really conform less than men when norms are not overtly prescribed, as Labov states in his formulation of the paradox. Rather, the sociolinguistic findings, as I understand them, only show that females' innovations are more successful than males'. Why is this so? Well, perhaps it is as simple as that: women dominate language, they have the power. Labov's gender paradox then simply describes two sides of the same coin of female dominance over language.

Linguistic change *from above* is based on prescription, sometimes maybe simply fashion (e.g. frequent use of loan vocabulary from a popular foreign language) newly adopted by the higher social classes. This type of change is rather exceptional. The usual case is change *from below*.

As Labov shows, the typical scenario of change from below is the frequent use of the innovative forms by particular women from a central social class (principle 1), e.g. the upper working or lower middle class. Those women, the main innovators, are upwardly mobile in the social hierarchy, with a rich and heterogeneous network of social connections. One part of Labov's solution to the gender paradox is that those innovative women, the leaders of linguistic change, are not the same women who are the most strictly conservative when it comes to stable linguistic variables.

Change from below is usually initiated by women, i.e. women adopt the new form more frequently and more consistently than men. However, sociolinguistic research also uncovered cases of change from below which are dominated by men (Labov 2006, 284): "But the cases where men are in the lead form a small minority. Furthermore, the male-dominated changes are all relatively isolated shifts."

The debate about the reasons for women's lead in language change and their stronger conservatism with respect to standard language accompanies sociolinguistics from the early days on and remains being controversial to the present day. The dispute is about, for instance, whether it is a sex or a

[5] Just to give it a straight formulation: women's deviations are innovations, men's deviations are mistakes.

gender issue, and if it is a gender issue whether it is motivated by women's alleged inferiority in society or by something else.

I agree with Chambers (2009) that biological differences between the sexes cannot be ignored in an explanation of these findings, simply for the reason that they persist even under conditions where women and men lead very similar lives, like the middle class in Western societies: "MC [middle class, RV] women … do not lead more or less insular lives than MC men. Yet we consistently find the same linguistic differences between men and women … namely that women use fewer stigmatized and non-standard variants than men of the same social group in the same circumstances." (Chamber 2009, 144)

Chambers (2009, 148ff) adds a number of psychological and neuropsychological findings which according to him strongly suggest that well-established cognitive differences between men and women lead to different command of language with a general (though small) advantage for women. Still, this line of reasoning lacks an explanation as to why women use their skill the way they do: Why do women avoid non-standard variants rather than make exaggerated use of them? Furthermore, truly social facts like Labov's principle 1 show that a biological explanation would be incomplete anyway.

But I also absolutely agree with Chambers' criticism that many solely sociological efforts have the presupposition "… that women are somehow compensating for shortcomings." He then continues: "And yet, the linguistic behaviour they are attempting to explain is not by any criterion negative. Any objective observer would be perplexed by the discrepancy between the linguistic behaviour and these explanations." (Chambers 2009, 147)[6]

If it was true that men dominated over women in all fields of society, why should women be able to compensate for their disadvantages in the

[6] Labov (2006, 276f, 291) shows some reservations towards Chambers' proposal. But the two views are not completely incompatible, as the following quote from an earlier paper of Labov's shows: "The explanations offered differ primarily in their emphasis on cultural or expressive traits as opposed to the political or economic position of women. It is interesting to note that no sociolinguistic argument views this behaviour of women as a form of superiority or an advantage to them. However, this does emerge in the popular view that women speak better or more correctly than men do. In disadvantaged communities, sensitivity to exterior standards of correctness in language is associated with upward social mobility. In the inner city black community, female students show greater success than males in school and greater employability. The effects of Principle [2, RV] can hardly be seen as the cause but rather a symptom of an overall readiness and opportunity to take advantage of prevailing community norms." (Labov 1990, 214)

field of language in the first place? Why do those overarchingly dominant men spare out language? They shouldn't if they were able to, but maybe they aren't.

Thus, if we follow Chambers' view, then Labov's principles 2-4 describe how women make use of their superior verbal abilities. It doesn't yet answer the question why women make this particular use of their advantage. Here is my personal speculative take on this.

Linguistic change is driven by frequency of use. The most common and most natural case is change from below. It is dominated by women, though we also rarely find situations where men dominate such a change. Thus, females must be able to trigger a higher frequency for their innovations than men do.

How might this come about? Usually, such a situation occurs when a group has higher internal consistency: women as a group might behave more consistently than men, i.e. they pick up each other's innovations more quickly and more consistently than men's innovations are picked up. Labov's principles suggest that this indeed might be the case: higher conformity of women to the norms of standard language should be accompanied by less between-speaker variance of women, i.e. higher conformity to each other, than we find for men. Men are too disorganised to counter the female dominance over language.

A further aspect is the self-reinforcing nature of this mechanism. If it is practised in the way just described, language de facto functions as a marker of social roles and status for women, not necessarily for men. A particular stance towards language norms then becomes part of a gender-specific stereotype for women. People in general tend to meet expectations linked to their social status including gender roles and so women, like anybody, prefer to show the expected behaviour.

3.2 Experiment 1: Gender specific stereotypes about grammatical taboos among university students

The existence of such stereotypes is an empirical assumption that can be put to test. This brings us back to the topic of this paper. Grammatical taboos belong to those non-standard phenomena – stigmatised to varying degrees – that should be subject to the sociocultural and in particular the gender differences just discussed.

Most of the sociolinguistic studies that contributed to the sociolinguistic findings summarised in Labov's four principles discussed above are based on *production* data. This is somewhat natural as many of these studies focus on phonological change. But it also has certain limitations.

Sociological explanations of women's more conservative attitude against stigmatised forms often rely on the presumption that their use is connected with the idea that people – especially women – think that their use has negative social consequences. But with production-oriented methods, such assumptions cannot be verified. This task requires perception-oriented methods.

In experimental morphosyntax, acceptability rating experiments are a well established method which can be put to use for such effects. The results can be compared with those of non-stigmatised forms of which we know quite a bit. Stimulus material can be prepared with specific gender (e.g. auditory stimuli) such that effects of gender on the acceptability of sentences can be investigated on both sides of the communicative channel and their combination (male/female participants judging sentences from male/female speakers). An acceptability study of this kind will be presented in section 4.

This current section presents the results of a pilot study that was designed to check whether the gender effects reported in the sociolinguistic literature could be a source of the variance in judgements of grammatical taboos by university students that was observed in our previous study described in section 2.

Method

The experiment was carried out as a written questionnaire. Each questionnaire consisted of a text that participants were supposed to read and a number of questions that had to be answered about the text. In particular, the text (about half an A4 page long) was a dialogue between a radio moderator and a person calling in via phone, as it can frequently be experienced on German radio. For the compilation of these texts we used material that can be found on the internet and adjusted it to the needs of the experiment. Moderator and caller in these dialogues discussed popular issues like nutrition, holidays, education, and consumer electronics.

We construed four different variants of such dialogues. The experiment tested four different conditions in a 2×2 Latin square design. The factors manipulated in the dialogue texts were the *gender* of the caller (male, female) and the inclusion/exclusion of stigmatised language. In particular, texts with stigmatised language contained one instance of the four grammatical taboo phenomena explored in (Vogel to appear) and in addition an English loanword – such so-called "anglicisms" are also stigmatised in the current popular discourse about language. Each of the

four lexical variants was used in each of the four conditions. We overall had 16 different variants of the questionnaire.

Each questionnaire contained only one text, such that only one of the four experiment conditions was tested with each participant. After reading the text, participants were asked to rate the caller according to several personal attributes. Participants were further asked to provide some personal information from which only their own gender is of interest here.

Two attributes ("social status" and "occupational status" of the caller) were elicited in the form of four-point Likert scales with the values "very high", "high", "low" and "very low". For a further list of attributes, participants were asked to mark them if they ascribe them to the caller. These attributes were "pleasant", "not pleasant", "nice", "arrogant", "educated", "uneducated", "naïve", and "unfriendly".

The goal of this pilot study is to figure out to what extent use of stigmatised expressions is associated with *negative stereotypes*, which kinds of negative connotations these are in particular and whether gender (of speaker, listener or particular combinations e.g. women listening to women) has an impact on the strength of these stereotypes.

Results

The participants of the experiment were 183 first year students of German studies at the faculty of linguistics and literature studies at the University of Bielefeld (132 female, 31 male), 149 of which were native speakers of German.[7] The distribution of the participants over the four experiment condition was as displayed in Table 3.

	male caller	female caller
without taboos	45 (12)	47 (12)
with taboos	44 (15)	47 (12)

Table 3: Distribution of the 183 participants over the four experiment conditions, male participants in brackets

I will now present the results for each of the elicited attributes. We start with the attribute "*occupational status*". As reported above, this attribute

[7] The results have been back-checked with the group of native speakers in isolation. As the observed effects were the same, there was no reason to exclude non-native speakers. So the results presented here include the data from all 183 participants.

was elicited in the form of a four-point rating scale with the labels "very high", "high", "low", "very low". Given the absence of a neutral scale centre, we expect "high" as the attribute that is more likely to be chosen under a neutral rating, assuming that participants hesitate to ascribe a negative attribute without having a reason for it.

The extreme points of the scale were chosen very rarely (four times each) so that the four point scale could be reduced to the two values "high" and "low", now including "very high" and "very low", respectively.

Our hypothesis then is that callers using grammatical taboos are more likely to be ascribed low occupational status. If there is a gender effect, this negative impact should be stronger for women than for men. 182 valid answers were given. Their distributions are displayed in Tables 4 and 5.

	high	low
with taboos	50.55% (46)	49.45% (45)
w/o taboos	72.53% (66)	27.47% (25)

Table 4: Attribution of high or low occupational status, depending on use of grammatical taboos, in %, exp. 1 (absolute counts in brackets). Test statistics: $\chi^2 = 8.3804$, df = 1, p-value = 0.003793 **; Cramer's V = 0.215 (small)

Table 4 shows the overall results for the use of taboo phenomena. Without using taboos, about three quarters of the participants gave a rating of high occupational status. This is expected under the assumption that this is the neutral value in this setting. With the use of taboos, we have equal chance for the ascription of high or low occupational status. Table 5 shows whether this affects both genders alike.

	male	female
with taboos	31.82%	65.96%
w/o taboos	28.89%	26.09%

Table 5: Attribution of low occupational status to male and female callers, depending on use of grammatical taboos, in %, exp. 1. Test statistics: $\chi^2 = 20.552$, df = 3, p-value = 0.0001304 *; Cramer's V = 0.336 (medium)**

Table 5 shows that male and female callers are rated quite differently. In fact, for female callers the attribution of low occupational status rises drastically, from about 26% to about 66%, whereas the minimal change from 29% to 32% for male callers is negligible. That is, for this attribute the negative effect of using taboo phenomena *is only there for female callers*, not for male callers, and it is quite large. The effect size measure, Cramer's V, is at medium effect size level, indicating that this is something that "careful observers" should be able to detect with their "naked eye" (Cohen 1992, 156).

The next attribute we study is the "social status" of the caller. This is somewhat different, though certainly not independent from occupational status. Subjects had to rate the rank of the caller in the social hierarchy. This attribute was elicited with the same four point Likert scale as before, and again, the extreme values had been chosen very rarely (5 times "very high", 2 times "very low"), so that the scale again was reduced to two values, "high" and "low". We received 181 valid answers. The distribution of ratings depending on the use of grammatical taboos overall is as given in Table 6.

	high	low
with taboos	52.75% (48)	47.25% (43)
w/o taboos	82.22% (74)	17.78% (16)

Table 6: Attribution of high or low social status, depending on use of grammatical taboos, in %, exp. 1 (absolute counts in brackets). Test statistics: $\chi^2 = 16.576$, df = 1, p-value = 4.675e-05 *; Cramer's V = 0.303 (medium)**

Participants gave a high rating even more consistently than for the occupational status for callers not using taboos. Those using taboos were given a low rating in about 50% of the cases, as before. Overall, the effect is larger than before with Cramer's V now marginally within the range of medium effect sizes. As we see in Table 7, this is due to degradation for the male caller that goes along with grammatical taboos, but still, it is less severe for males than for females.

	male	female
with taboos	38.64%	55.32%
w/o taboos	15.91%	19.57%

Table 7: Attribution of low social status to male and female callers, depending on use of grammatical taboos, in %, exp. 1.
Test statistics: χ^2= 20.907, df = 3, p-value = 0.00011 *;**
Cramer's V = 0.340 (medium)

We now turn to those attributes that were ascribed in an unscaled way. We start with the two attributes "educated" and "uneducated". These have been elicited independently, but as they are clearly related, we view them as one attribute. No participant rated the caller as educated and uneducated at the same time. So there were three possible outcomes: callers were rated as educated, uneducated or neither. The three outcomes can be seen as forming a gradient scale of "educatedness". The results are listed in Table 8.

	educated	neither	uneducated
With taboos	39.6% (36)	28.6% (26)	31.9% (29)
w/o taboos	75.0% (69)	20.7% (19)	4.3% (4)

Table 8: Ascription of the attributes educated, uneducated or neither, depending on use of grammatical taboos, in %, exp. 1 (absolute counts in brackets)
Test statistics: χ^2 = 30.395, df = 2, p-value = 2.511e-07 *;**
Cramer's V = 0.408 (medium)

We see that without use of taboo expressions, callers were rated as educated by about three quarters of the participants, and in only about four per cent of the cases they were rated as uneducated. With the use of taboo expressions, the three outcomes are chosen nearly evenly, with still a slight advantage for "educated". Similarly to what we saw before with the ascription of social status, the effect is larger for female than for male callers, as shown in Table 9.

		educated	neither	uneducated
male caller	with taboos	54.6% (24)	13.6% (6)	31.8% (14)
female caller	with taboos	25.5% (12)	42.6% (20)	31.9% (15)
male caller	w/o taboos	77.8% (35)	20.0% (9)	2.2% (1)
female caller	w/o taboos	72.3% (34)	21.3% (10)	6.4% (3)

Table 9: Ascription of the attributes educated, uneducated or neither, depending on use of grammatical taboos, in %, exp. 1 (absolute counts in brackets)
Test statistics: $\chi^2 = 41.811$, df = 6, p-value = 2.003e-07 *;**
Cramer's V = 0.338 (medium)

The overall impression is that females using taboo expressions run a much higher risk of not appearing as educated as men do – though use of taboos is not without risk for men, either. This again fits into the sociolinguistic picture discussed in section 3.1.

The other tested attributes express positive or negative subjective impressions. Most of them showed no particular effects. Table 10 summarises their results, for completeness sake.

Two of these subjective attributes did show interesting contrasts. One of them is the attribute "pleasant" ("sympathisch"). Table 11 displays the results now differentiated for male and female experiment participants.

	Nice ("nett")	not pleasant ("unsympathisch")	arrogant ("arrogant")	Unfriendly ("unfreundlich")
with taboos	74.4% (68/91)	7.7% (7/91)	4.4% (4/91)	0% (0/91)
w/o taboos	77.2% (71/92)	5.4% (5/92)	7.6% (7/92)	0% (0/92)

Table 10: Results for several subjective attributes depending on use of taboo expressions, exp. 1; no relevant contrasts; in brackets: counts of "yes"-answers/totals

		male partic.	female partic.
male caller	with taboos	20.0% (3/15)	72.4% (21/29)
female caller	with taboos	75.0% (9/12)	69.7% (23/33)
male caller	w/o taboos	58.3% (7/12)	74.3% (26/35)
female caller	w/o taboos	41.7% (5/12)	68.6% (24/35)

Table 11: Results for the attribute pleasant ("sympathisch") by gender of experiment participants, depending on use of taboos and gender of the caller; in brackets: counts of "yes"-answers/totals Test statistics: $\chi^2 = 19.377$, df = 7, p-value = 0.007086 **; Cramer's V = 0.325 (medium) (male partic. only): $\chi^2 = 8.9211$, df = 3, p-value = 0.03036 *; Cramer's V = 0.418 (medium)

We see in Table 11 that male and female participants use the label "pleasant" quite differently. Whereas female participants do not seem to make a difference with respect to the gender of the caller and their use of taboo expressions with values around 70 % in each of the four cases, male participants seem to make huge differences, such that male callers using taboos receive the lowest score of 20 % whereas female callers using taboos recieve the highest score of 75 %. Callers not using taboos are in the middle. But note the quite low number of male experiment participants, so the gender difference we observe still has to be handled with some care.

The finding, here for male participants only, that callers using taboos with the own gender of the participant are rated more negatively than callers of the other gender can also be found with the final attribute that we inspect, "naïve" ("naiv"), see Table 12.

	gender	
	own	other
with taboos	56.0% (28/50)	39.0% (16/41)
w/o taboos	25.5% (12/47)	31.1% (14/45)

Table 12: Results for the attribute naïve ("naiv"), depending on use of taboos and gender of the caller (own vs. other gender of experiment participant); in brackets: counts of "yes"-answers/totals
Test statistics: $\chi^2 = 10.869$, df = 3, p-value = 0.01245 *;
Cramer's V = 0.244 (small)

According to the test statistics of Table 12, this is a significant, but small effect suggesting that subjects are a bit stricter against use of taboos by members of their own gender.

But this evidence is large enough for us to be cautious when preparing experiments with auditory stimuli: there might be effects which are solely caused by the interaction of the gender of the person who produced the stimuli and the gender of an experiment participant. That this is indeed worthwhile, will be shown in the experiment presented in the next section.

First, a brief summary of this section:

1. Sociolinguistic evidence points to the leading role of women both in preserving language norms and in establishing new ones. The most frequent kind of change, change from below, originates in a central social group (lower middle class). Its most active members are upwardly mobile women.
2. In the German-speaking countries, education is presumably the most important vehicle of upwards mobility. University students, like the participants of the experiments presented in this paper, are therefore quite an interesting population to study under the here chosen sociolinguistic focus, both in order to put to test sociolinguistic hypotheses, and with respect to the linguistic variation that emerges within this important group.
3. Very much in accordance with the sociolinguistic wisdom, experiment 1 showed that our university students' negative evaluation of using taboos is more severe for females than for males, especially for important attributes like occupational status, social status and education – these negative ascriptions are given by male and female participants at the same rate. When women

conform more strictly to standard norms, thus, they also *fulfill language-related expectations* connected with their gender roles. This is not to suggest an alternative explanation for the role of women in language change as described by Labov's four principles. Rather, it seems natural to me that behaviour (in the past and present) and expectations (about present and future behaviour) reinforce each other.

4. Experiment 2: Gender and sociocultural background as determinants of the acceptability of grammatical taboos

Experiment 1 explored a bundle of five taboo phenomena occurring in one text in order to determine whether core insights of sociolinguistic theory manifest themselves also with these phenomena among university students. We received a robust positive answer. However, because the five phenomena have been explored together, the extent of the contribution of each of the five phenomena is unclear. Our earlier study already corroborated that anglicisms are much less severely stigmatised than the four grammatical taboo phenomena, and our initial assumption that two of the tested grammatical taboos are clearly stronger than the other two has also been confirmed.

This suggests that the five phenomena contribute with varying degrees to the findings in experiment 1. To test this, we need to inspect each phenomenon separately. This has been carried out in experiment 2 which I will introduce now.

Method

This follow-up experiment to the questionnaire study in (Vogel submitted) was again an acceptability judgement experiment, but with a number of changes. First, only the linguistically most realistic judgement type P ("Is this a possible sentence of German?") has been used. Second, auditory stimulus material was used, and third, as a consequence of this, immediate reactions to a presented stimulus were elicited which required a simplification of the judgement task, which is now a binary yes/no judgement.

In order to test systematically for the effect of gender, stimulus material has been produced in two versions, one with a female and another with a male speaker. Single participants rated the stimuli from either female or the male speaker. Stimuli with different gender were not used within experiment sessions.

We also controlled for gender as a factor on the side of experiment participants, so that each gender was represented to the same amount. Furthermore, we widened our view to include students from faculties other than our own faculty of linguistics and literature studies (LiLi) at the University of Bielefeld.

Each of these three factors was balanced so that the 80 participants of the experiment are divided into eight groups of ten participants each (LiLi/other faculty × male/female participant × male/female stimuli). This allows for an investigation of the factor gender from three different angles (gender of stimulus, participant and (non-)identity between the two).

Participants were asked to provide further information about themselves, respecting anonymity. As I briefly discussed above, university students are an interesting group, because they often belong to the group of upwardly mobile people who play a crucial role in language change. However, only a subgroup of them actually does so, namely those whose parents themselves are not already in the upper classes. The indicators for this that we use in this study are the highest educational levels that each of their parents has reached.

In sociological studies, the high school grade, called "Abitur" in Germany, has been identified as the grade which makes an important sociocultural difference for many dimensions of everyday life. It usually serves as a ticket to a solid middle class life. According to their specifications, participants have been divided into three groups, depending on whether none, only one or both of their parents have the Abitur. This factor could not be balanced, of course, so it is distributed unevenly among the eight groups. Table 13 gives an overview.

		parents: Abitur		
Faculty	gender (partic.)	one	both	neither
LiLi	m	4	9	7
other	m	7	4	9
LiLi	w	6	7	7
other	w	5	7	8
total		22	27	31

Table 13: Distribution of participants according to educational level of their parents, gender and faculty membership in exp. 2

I assume that those participants whose parents both have the Abitur are not or only minimally upwardly mobile. Those neither of whose parents have the Abitur clearly have already climbed up the ladder, with the university being a world their parents had no experience with and can hardly help them through.

Both groups are homogeneous in the educational background of their parents. This distinguishes them from the group where only one of the parents has the Abitur. In such a situation, we might find a quite heterogeneous network including working and middle class members in the broader family. Homogeneous educational background of the parents makes such a background a bit less likely.

Of course, such interpretations appear quite far-fetching. On the other hand, we will see below that they are not unjustified. We expect different attitudes towards stigmatised speech to follow from this parameter. Participants from the "neither" group might have the highest reservations towards stigmatised speech, eager not to reveal their non-academic, perhaps non-middle class roots. The "both" group might not experience such fears and show a bit more tolerance, but it presumably sticks conservatively to the standard language. The "one" group, finally, might on the one hand experience less fear than the "neither" group, and on the other hand be less conservative and more tolerant than the "both" group due to their more heterogeneous sociocultural background.

Pressure towards linguistic conformity is therefore presumably highest in the "neither" group and lowest in the "one" group. Please also note that the "one" group is the smallest group in our random sample (27.5%) which suggests that homogeneity in educational and sociocultural background may be the preferred option. The "neither" group makes up 38.75% of the sample whereas the "both" group is one third (33.75%). Thus, roughly between one half and two thirds of the participants (and perhaps of our university students) might be upwardly mobile (with varying degrees).

The experiment was run using the *DMASTR* software (*DMDX*).[8] Test stimuli were presented in randomised order to the participants, via ear phone. After each test item, participants were given a visual signal on the computer screen to press one of two buttons indicating whether they considered the sentence they heard as a possible German sentence or not. The results were analysed statistically with the software package *R* (R Core Team 2016).

[8] It was developed at Monash University and at the University of Arizona by K.I. Forster and J.C. Forster.

Material

The experiment tested the four grammatical taboo phenomena that have already been investigated in (Vogel to appear), see (4) and (5). The test sentences have been reused with only minor changes where this seemed necessary. For each of the four phenomena eight lexical variants were construed and used in a 2×2 latin square design with the factors ±grammaticality and ± taboo compliance: the otherwise grammatical sentences with taboo violations are paired with a semantically equivalent syntactically unmarked variant without taboo violation. In addition, ungrammatical variants of these two structures are construed where the ungrammaticality is due to an inflection error on the finite verb. We thus have 32 test sentences for each of the four phenomena, overall 128 all of which have been included in every trial. In our discussion of the results below, only a subset of the test items will be considered.

The material further included 47 filler sentences out of which 10 were used as training sentences at the beginning of each session. These are not further taken into account. Among the 37 fillers available for closer inspection were 14 sentences that contained English loanwords (so-called "anglicisms") and were otherwise unmarked, 9 sentences that count as morphosyntactically marked and 12 ungrammatical sentences.[9] Most filler sentences were reused from the experiment in (Vogel to appear).

Two of the 80 participants were excluded from the analysis, because their positive judgements of the ungrammatical filler sentences were higher than 50%.

Results

Table 14 displays the proportions of positive responses for various sentence types in exp. 2 in comparison to the equivalent results from the questionnaire experiment in (Vogel submitted). In each case, identical item sets are compared.[10] Table 14 reports the results only for the group of 40 students from the LiLi faculty, which can be seen as comparable to the group in the questionnaire experiment.

[9] Two filler sentences had other special properties which are not of interest here.

[10] The unmarked sentences used for Table 14 are the unmarked grammatical test conditions for the four taboo phenomena, combined. Likewise, the compared ungrammatical sentences are the ungrammatical test conditions of the four taboo phenomena.

sentence type	mean rating (Vogel to app.)	prop. accept. exp. 2	Diff.	Cohen's h
unmarked	0.929	0.980	0.051	0.256 (small)
anglicism	0.788	0.901	0.113	0.317 (small)
marked	0.392	0.506	0.114	0.230 (small)
ungrammatical	0.090	0.020	0.070	0.326 (small)
gr. taboos (combined)	0.361	0.436	0.075	0.153 (–)
aux. *tun*	0.254	0.299	0.045	0.101 (–)
weil V2 clause	0.375	0.375	0.000	0.000 (–)
d-pronoun	0.396	0.505	0.109	0.220 (small)
double perfect	0.397	0.566	0.169	0.340 (small)

Table 14: Comparison of mean ratings in (Vogel submitted) and proportions of positive replies in experiment 2 of the present study, with absolute differences and the Cohen's h effect size measure; only results of students from the LiLi faculty are included for exp. 2 here

Table 14 gives an impression of the impact of the different elicitation methods used in the two studies: written questionnaire with 7-point rating scale (Vogel submitted) vs. auditory stimulus presentation and binary judgement (exp. 2). As effect size measure, Cohen's h is used (Cohen 1988, ch. 6) which measures the effect size of differences between proportions.[11]

Overall, the ratings in exp. 2 are a bit more positive than for the first study. While this improvement is in the range of a small effect size for most sentence types, for the four grammatical taboos together it is less than small. Inspection of the individual phenomena shows that this is due to the *salient taboos* auxiliary *tun* and *weil* V2-clause, whereas the non-

[11] For this calculation, a proportion P is rescaled as $\varphi = 2 \arcsin \sqrt{P}$. The (non-directional) effect size h for the contrast between two proportions P_1 and P_2 then is $|\varphi_1 - \varphi_2|$. An $h > 0.2$ counts as small effect, $h > 0.5$ as medium and $h > 0.8$ as large effect. Note that the values from the rating study in (Vogel to appear) here are treated as approximating proportional data.

salient taboos *d*-pronoun and double perfect show small effect size like the ordinary sentence types.

Let us now turn to the sociolinguistic factors inspected in experiment 2. We will first look at the main factors in isolation. Table 15 displays the contrasts for the factor faculty membership.

	faculty					
	LiLi	other	diff.	Cohen's h		p-value
combined	0.436	0.300	0.136	0.282	(small)	5.366e-12 ***
aux. *tun*	0.299	0.184	0.115	0.271	(small)	0.00148 **
weil V2-clause	0.375	0.160	0.215	0.494	(≈medium)	3.598e-09 ***
d-pronoun	0.505	0.342	0.162	0.330	(small)	7e-05 ***
double perfect	0.566	0.512	0.054	0.109	(–)	0.2048

Table 15: Splitted results, according to faculty membership, for grammatical taboos combined and by phenomenon; proportions of positive responses; Cohen's h effect size and p-value of test for equality of proportions

We see that in general students from other faculties are more severe than students from the faculty of linguistics and literature studies. The effect is largest, of nearly medium size, for *weil* V2-clauses and smallest, below the small effect size threshold, for the double perfect.

This suggests that students from the LiLi faculty are more tolerant towards stigmatised constructions, which is most likely a professional bias that is working against stigmatisation to some degree. The fact that the contrast is nearly absent for the double perfect might indicate that this phenomenon suffers the weakest stigmatisation (which the absolute acceptability value also shows). In Table 16, the contrasts for male and female participants are displayed.

	gender						
	male	female	diff.	Cohen's h		p-value	
combined	0.387	0.354	0.033	0.068	(–)	0.1011	
aux. *tun*	0.294	0.195	0.099	0.232	(small)	0.006315	**
weil V2-clause	0.245	0.294	0.049	0.111	(–)	0.1974	
d-pronoun	0.489	0.364	0.125	0.253	(small)	0.002396	**
double perfect	0.519	0.560	0.041	0.082	(–)	0.3505	

Table 16: Splitted results, according to the gender of the participants, for grammatical taboos combined and by phenomenon, proportions of positive responses; Cohen's h effect size and p-value of test for equality of proportions

While female participants have higher judgements for the double perfect and *weil* V2-clauses, the opposite is true of auxiliary *tun* and *d*-pronouns. Only in the latter two cases the level of small effect size is reached. The overall tendency is in line with our hypotheses of women conforming more to the standard than men, but its rather small amount might be surprising, as well as the fact that it seems to be restricted to auxiliary *tun* and *d*-pronouns. In Table 17, the results for the gender difference in the stimuli are presented.

	gender in stimuli						
	male st.	female st.	diff.	Cohen's h		p-value	
Combined	0.396	0.347	0.049	0.101	(–)	0.01418	*
Aux. *tun*	0.227	0.260	0.033	0.077	(–)	0.3948	
Weil V2-clause	0.339	0.204	0.135	0.306	(small)	0.0002328	***
d-pronoun	0.490	0.367	0.123	0.249	(small)	0.002815	**
double perfect	0.520	0.559	0.039	0.078	(–)	0.3776	

Table 17: Splitted results, according to gender of stimulus material, for grammatical taboos combined and by phenomenon; proportions of positive responses; Cohen's h effect size and p-value of test for equality of proportions

We can make similar observations as before: male and female stimuli each have better ratings in two of the four cases, whereby only those cases where ratings for male stimuli are higher reach small effect size. The cases are not the same as before, though: while *d*-pronouns have higher ratings from men and with male stimuli, auxiliary *tun* has higher ratings from men and for the female stimuli. Similarly, *weil* V2-clauses have higher ratings from women and for male stimuli. The double perfect has higher ratings from women and for female stimuli, but in both cases with less than small effect size.

These observations call for a closer inspection of the interaction of the two gender factors. We will approach this in two steps. First, we look at the binary factor of participants judging their own or the other gender. This is displayed in Table 18.

This factor shows a consistent picture of participants judging material from the other gender better than material from their own gender. The effect has small size for the two non-salient taboo phenomena, and nearly small size for the salient taboos.

What we cannot judge from Table 18 is whether both genders make a difference between their own and the other gender in the same way and whether this is constant between the four taboo phenomena. Table 19 unfolds this.

	gender in stimuli						
	own	other	diff.	Cohen's h		p-value	
combined	0.320	0.423	0.103	0.214	(small)	1.533e-07	***
aux. *tun*	0.205	0.284	0.079	0.184	(–)	0.03098	*
weil V2-clause	0.229	0.331	0.082	0.186	(–)	0.02697	*
d-pronoun	0.356	0.498	0.142	0.289	(small)	0.0005039	***
double perfect	0.485	0.595	0.109	0.220	(small)	0.008312	**

Table 18: Splitted results, according to participants judging material with their own or the other gender, for grammatical taboos combined and by phenomenon, proportions of positive responses; Cohen's h effect size and p-value of test for equality of proportions

gender of partic.	male		female			
gender in stimulus	own	other	own	other	p-value	
combined	0.360	0.413	0.282	0.433	5.974e-08	***
aux. *tun*	0.238	0.344	0.176	0.216	0.004204	**
weil V2-clause	0.272	0.219	0.189	0.407	8.007e-05	***
d-pronoun	0.480	0.497	0.239	0.500	8.465e-07	***
double perfect	0.444	0.592	0.525	0.597	0.02359	*

Table 19: Splitted results, by gender of participants and own/other gender of the stimulus, for grammatical taboos combined and by phenomenon; proportions of positive responses; p-value of test for equality of proportions

Several things are remarkable about Table 19. First, for each row we can identify one cell whose value differs from the three others which are quite close. Those exceptional cells are highlighted. This clearly shows that we are dealing with *interactions* of gender-related factors. Second, for the two salient phenomena, the exceptional cells diverge by exceptionally higher values, whereas for the non-salient phenomena they diverge by lower values. Third, for each of the four phenomena, this exceptional cell is in a different column.

Estimation of effect size is a bit more complex because we now have to take into account multiple pairwise comparisons. In Table 20, the values for Cohen's h are given for the three pairwise comparisons of the exceptional cells that we identified in Table 19 with the three other cells. The final column contains the means of those three values which is our approximation of the effect size associated with these particular cells.

gender of partic.	male		female			
gender in stimulus	own	other	own	other	mean h	
combined	0.167	0.276	—	0.317	0.253	(small)
aux. *tun*	0.234	—	0.388	0.287	0.303	(small)
weil V2-clause	0.286	0.410	0.484	—	0.393	(small)
d-pronoun	0.509	0.543	—	0.549	0.534	(medium)
double perfect	—	0.297	0.162	0.307	0.256	(small)

Table 20: Cohen's h for pairwise comparisons with the highlighted values in Table 19, by row, plus means for each row

The effect is largest for the *d*-pronouns. With even medium effect size, it is surprisingly high. For the other non-salient phenomenon, double perfect, it is at the lower edge of a small effect, however, and perhaps negligible. For auxiliary *tun* and *weil* V2-clause, we observe small effect sizes.

How can we interpret these results? Obviously, male and female participants differ in their attitude towards each of the four phenomena, and they do so in different ways each time.

With respect to auxiliary *tun*, male participants have a certain tolerance if it is used by women. Female participants make little difference as to the gender in the stimuli. What does this signal? The most plausible explanation to me is that if a stigmatisation is accepted by speakers, then the gender in the stimuli makes no difference and we expect low values, as is the case with female participants. Male participants, then, know about the stigmatisation, but accept it to a lesser degree. Therefore, they show more tolerance towards female speakers using auxiliary *tun*, but cannot show the same tolerance towards their own gender (which they identify with) due to the risks of social stigmatisation. In other words, men are aware of the social stigmation associated with using the construction, but unwilling to exert it themselves against the other gender.

With *weil* V2-clauses, we have the same situation, but with gender roles switched: this time, it is the female participants who show a readiness to accept the construction, but again, because of the stigmatisation they show this only towards the other gender. The pattern for *weil* V2-clauses is closest to what we observe with ordinary markedness. When we split the results for ordinary markedness presented in Table 14 in the same way as in Table 19, we get the picture in Table 21.

	gender in stimulus	
partic.	own	other
male	0.500	0.535
female	0.505	0.680

Table 21: Splitted results for marked filler sentences, exp. 2, participants from the LiLi faculty only, proportions of positive replies

The non-salient phenomena are candidates for what has been characterised as "change from below" by Labov (1990, 2006). We have learned from Labov that such changes are usually driven by one gender, mostly by women, but rarely also by men. If this were the case here, then we would expect the construction to be accepted by the active gender to a higher degree independent of the gender in the stimulus and by members of the passive gender for their own gender to a lower degree than for the other gender.

And this is indeed what we observe here. For *d*-pronouns, acceptability is similarly high for male participants and male stimuli, but female participants degrade the construction if they hear it from a female voice. Thus, *d*-pronoun is a candidate for change from below driven by males. The double perfect, on the other hand, shows again the mirror image and thus is a candidate for change from below driven by females.

In the latter case, we also expect effects of social class, whereas changes driven by males are more diffuse in this respect. We can check for this by inspecting the factor of the educational background of the parents. The results are given in Table 22.

	parents' Abitur						
	one	both	none	Cohen's h		p-value	
combined	0.447	0.324	0.354	0.169	(–)	2.119e-06	***
aux. *tun*	0.407	0.230	0.135	0.420	(small)	2.565e-09	***
weil V2-clause	0.295	0.284	0.238	0.086	(–)	0.3699	
d-pronoun	0.391	0.405	0.472	0.110	(–)	0.194	
double perfect	0.699	0.376	0.565	0.441	(small)	1.514e-09	***

Table 22: Splitted results, by educational background of participants' parents, for grammatical taboos combined and by phenomenon, proportions of positive responses; Cohen's h effect size and p-value of test for equality of proportions

The relative contrasts that we expect from our considerations about this category are matched to some extent by the results for the taboos combined, insofar as the group with only one parent with Abitur turns out to be the most tolerant, due to their presumably more heterogeneous sociocultural family background. The only phenomenon where this does not hold is the *d*-pronoun. The "none" group has the lowest values for the two salient phenomena, but the highest value for *d*-pronouns and a still high value for the double perfect.

This confirms our suspicion that the two non-salient phenomena can be interpreted as cases of "change from below". From this perspective, we also expect the finding that the "both" group has the lowest ratings for the double perfect.

Three remarkable "outliers" are highlighted in Table 22. Whether these hold irrespective of gender will be inspected next. We therefore inspect the interaction of parents' educational background with participants' gender. We first take a look at the salient phenomena (see Table 23).

		parents' Abitur						
taboo	partic.	one	both	none	Cohen's h		p-value	
aux. *tun*	male	0.558	0.186	0.191	0.530	(medium)	1.625e-09	***
	female	0.256	0.276	0.079	0.357	(small)	0.0003569	***
weil V2-cl.	male	0.307	0.188	0.248	0.185	(–)	0.1662	
	female	0.284	0.379	0.229	0.219	(small)	0.04964	*

Table 23: Splitted results, by gender of participants and educational background of their parents, for salient grammatical taboos; proportions of positive responses; mean Cohen's h for three pairwise comparisons, p-values of test for equality of proportions

We see for auxiliary *tun* an extraordinarily high value for the male participants of the "one" group. Their acceptability rating goes up to 55%. This is in line with the sociolinguistic finding of men following the standard less consistently than women. Apart from this, it is noteworthy that this is restricted to males from the "one" group which we expect to be the most tolerant one. The lowest value is for female participants of the "none" group. Auxiliary *tun* is the strongest taboo inspected here and women from the "none" group are those who have to fear stigmatisation the most. Both observations are in line with what we have learned so far.

With respect to *weil* V2-clauses, contrasts are smaller overall, but we see that the females from the "both" group differ at about 20% from males of this group. Given our observation that this is the weaker of the two salient taboos, we can interpret the results such that there might be a *de-stigmatisation* of this construction being underway, which would then be an instance of change from above with females from the highest group (the "both" group) as initiators.

		parents' Abitur						
taboo	partic.	one	Both	none	Cohen's h		p-value	
d-pron.	male	0.395	0.447	0.593	0.266	(small)	0.01174	*
	female	0.386	0.363	0.348	0.053	(–)	0.8518	
dbl. perf.	male	0.767	0.294	0.533	0.659	(medium)	7.476e-10	***
	female	0.632	0.456	0.598	0.070	(–)	0.02973	*

Table 24: Splitted results, by gender of participants and educational background of their parents, for salient grammatical taboos, proportions of positive responses; mean Cohen's h for three pairwise comparisons, p-values of test for equality of proportions

For the non-salient taboos, our conclusions drawn above are also confirmed. The *d*-pronouns receive a remarkably high score from male participants with the lowest educational background. This is in line with our idea that this phenomenon instantiates male-driven change from below. With respect to the double perfect, we see a remarkably low score for males from the socioculturally highest "both" group. Again, this is in line with our idea that the double perfect is change from below driven by females.

The conclusions drawn from these inspections have been put to further testing with more advanced methods of inferential statistics. This is especially important, as we are dealing with the interaction of several factors and the factor "educational background" is unbalanced.

Such calculations can be carried out with generalized linear mixed models. We used the software package *lme4* for R (Bates et al. 2015) for these calculations. GLMMs use a variety of fixed and random factors to *model* the distribution in a given data set. χ^2 goodness of fit tests are carried out to compare different models, in particular models with higher or lower complexity. What is tested is whether enriching the model by the inclusion of further factors leads to a significant improvement of the model. By repeated application of model testing an optimal model can be identified. This procedure has been carried out for each of our four taboo phenomena. Table 25 lists the fixed factor model and the results of the

goodness of fit tests as compared to a model with only the random factors.[12]

	fixed factors	χ^2	df	p-value	
aux. *tun*	EDU * GENDER * OWNG	38.011	11	7.79e-05	***
weil V2-clause	GROUP	8.3915	1	0.00377	**
d-pronoun	GROUP + OWNG	7.1533	2	0.02797	*
double perfect	EDU	9.827	2	0.007347	**

Table 25: Optimal generalised linear mixed effects models for each taboo phenomenon: fixed factor specifications and results of goodness of fit tests against the null model with only random factors

The results of these tests by and large confirm what we saw from the previous inspections. This modelling approach additionally reduces the various factors we observed to take effect in our four taboo phenomena to a minimal set that is needed for an optimal data fit.

Auxiliary *tun* turns out to be a paradigm case for the interaction of gender and sociocultural status that is typical of strong sociolinguistic effects in standard languages – its stigmatisation appears to be quite stable. The other phenomena are less typical, and only some of the observed factors are relevant. The double perfect can be identified as instance of change from below due to the strong effect of educational background. The *d*-pronoun case shows the gender effects without the effect of educational background. We already identified it as an instance of change from below initiated by males. For the *weil* V2-clause we saw a larger effect of faculty membership and some smaller effects that do not provide significant improvements of the linear model.

5. Conclusion

Starting point for the research presented in this paper was the observation in (Vogel to appear) that the rating of grammatical taboos by university

[12] The abbreviations used in Table 25 have the following meanings: GROUP: faculty of participants (LiLi or other); GENDER: gender of participants; OWNG: gender in stimulus matches (or not) the gender of participants (own or other); EDU: educational background of the parents (Abitur for none, one or both).

students is in the range of morphosyntactic markedness, but differs from the latter by a larger between-subject variance.

Here we focused on the question, whether factors that have been established by the sociolinguistic research in the previous decades, in particular gender in interaction with socioeconomic class and sociocultural differentiation, can account for the variances in the rating of grammatical taboos within such a homogeneous group like university students.

We have not only been able to confirm this, but we also found out that each of the four investigated grammatical taboos of German has its own characteristic sociolinguistic profile. Gender plays some role in the four cases, but quite differently in each case (Table 19). We have taken the educational background of participants' parents as indicator of different sociocultural and socioeconomic status. Coarse-grained as this might be, we nevertheless observed interesting effects of this factor that fit quite well with the idea that our two non-salient taboos are instances of change from below in the sense of Labov (1990, 2006). Likewise, the stable and salient taboo over auxiliary *tun* emerged as a typical example for the sociolinguistic differentiation of a linguistic variable.

In addition, we observed that students of a faculty of languages and literature studies seem to have acquired already a certain tolerance towards "deviant" language as part of their professional training.

Still, the conclusions that we have drawn on the basis of our data need to be taken as provisional. Studies like this are rare. Replication studies are certainly in order to base our findings on more solid ground.

References

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Chambers, J. K. (2009). *Sociolinguistic theory* (rev. ed.), *Language in society*. Chichester, vol. 22, West Sussex [a.o.]: Wiley-Blackwell.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. London: Routledge.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdayle, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin 112*(1), 155–159.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change 2*(2), 205–254.

Labov, W. (2006). *Principles of Linguistic Change. Volume 2: Social Factors*. Oxford, Malden, Mass.: Blackwell Publishers.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romano, J., J.D. Kromrey, J. Coraggio, and J. Skowronek (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys? In: Annual meeting of the Florida Association for Institutional Research, 2006, Cocoa Beach, Florida, USA.

Vogel, R. (to appear). Grammatical Taboos. An investigation on the impact of prescription in acceptability judgement experiments. To appear in Zeitschrift für Sprachwissenschaft.

von Polenz, P. (2000). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart* (2nd, revised edition ed.), Volume I: Einführung, Grundbegriffe, 14.-16. Jahrhundert. Berlin: de Gruyter.

Weiß, H. (2004). Zum linguistischen Status von Standardsprachen. In M. Kozianka, R. Lühr, and S. Zeilfelder (Eds.), *Indogermanistik – Germanistik – Linguistik. Akten der Arbeitstagung der Indogermanistischen Gesellschaft, Jena, 18.-20.9.2002*, pp. 591–643. Hamburg: Verlag Dr. Kovač.

Weiß, H. (2005). Von den vier Lebensaltern einer Standardsprache. *Deutsche Sprache 2005*(4), 289–307.