Ralf Vogel*

Grammatical taboos

An investigation on the impact of prescription in acceptability judgement experiments

https://doi.org/10.1515/zfs-2019-0002

Abstract: This explorative study focuses on *grammatical taboos* in German, morphosyntactic constructions which are subject to stigmatisation, as they regularly occur in standard languages. They are subjected to systematic experimental testing in a questionnaire study with gradient rating scales on two salient and two non-salient grammatical taboo phenomena of German. The study is divided into three subexperiments with different judgement types, an aesthetic judgement, a norm-oriented judgement and the sort of possibility judgement that comes closest to linguists' understanding of grammar. Included in the investigated material are also examples of ordinary gradient grammaticality: unmarked, marked and ungrammatical sentences. The empirical characteristics of grammatical taboos are compared to those ordinary cases with the finding that they are rated at the level of markedness, but differ from ordinary markedness in that they produce a different pattern of between-subject variance. In addition, we find that grammatical taboos have a particular disadvantage under the aesthetic judgement type. The paper also introduces the concept of *empirical grammaticality* as a necessary theoretical cornerstone for empirical linguistics. Methodically, the study applies a mix of parametric and non-parametric methods of statistical analysis.

Keywords: experimental morpho-syntax, prescription, empirical grammaticality, morphosyntactic markedness

1 Introduction

The initial motivation for the study presented in this paper is a problem that everyone is facing who carries out empirical studies on grammaticality: the unknown influence of prescription and ideological bias on the outcomes of such studies. The participants of elicitation experiments, typically linguistically naïve, rarely understand the difference between the "natural" rules and constraints of their

***Corresponding author: Ralf Vogel,** Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft, Bielefeld, Germany, e-mail: ralf.vogel@uni-bielefeld.de

language's grammar (which linguists are interested in) and prescriptive constraints (which often are seen as uninteresting artefacts, not only by theoretical linguists).

The perhaps most widespread way of dealing with this issue is to see them as confounds. In designing experiments, we avoid any linguistic tokens in the stimulus material that are suspicious of triggering negative responses for other reasons than violations of the "natural" grammatical constraints we are interested in.

There are two related problems resulting from this practice. First, certain linguistic phenomena are excluded from empirical investigation and thereby are looming to remain out of sight of grammatical theory, with the risk of delivering a distorted picture of a language's grammar. Second, the stigmatisation of the avoided constructions is repeated and inadvertently enhanced – as if linguists implicitly accepted the restrictions imposed by prescription, taking the posited prescriptive constraints for granted without ever evaluating them empirically.

The latter consequence is at odds with the character of modern linguistics as an empirical discipline. No factual assumptions should be exempted from empirical testing in linguistics. On the contrary, if we presuppose that prescription has an influence on the outcomes of elicitation experiments, and we obviously do so, empirical linguists are obliged to investigate this factor both in its quantitative and in its qualitative impacts. These could be impacts on judgements only, which would reduce the task to the solely methodological question of optimising elicitation methods. But most researchers on the historical development of standard languages agree that prescription played a role in *shaping their language systems*, and thus assume prescription to be a necessary factor (though perhaps a smaller one) in an explanatory theory of grammar as it is.

The present study considers both dimensions of the problem. It is focusing on *grammatical taboos* in German, morphosyntactic constructions which are subject to stigmatisation. In a complex experimental setting using different instructions and judgement types, it is tested whether the influence of prescription on the rating of grammatical taboos in morphosyntactic experiments can be minimised. But beyond this, grammatical taboos are also explored as a *specific type of morphosyntactic markedness* the characteristics of which can be determined by means of morphosyntactic experimentation.

In order to do so, other categories of (gradient) grammaticality need to be determined in their empirical characteristics. This is carried out in the spirit of recent developments concerning statistical methodology in the behavioural sciences. Their consequences for experimental morphosyntax are discussed in Section 2. There, it is also argued for the concept of *empirical grammaticality* as a necessary ingredient of a consequently empiricist view of grammar.

Section 3 introduces grammatical taboos and explains their operationalisation for the experiment that is presented in Section 4. The paper concludes with a

summary in Section 5. A central role in the explanation of the results is played by a characteristic property of grammatical taboos that I term the *paradox of grammatical taboos*, which results from the fact that only existing constructions can be ruled out by taboo constraints.

2 General assumptions

This section introduces basic assumptions on the language system, its empirical exploration with acceptability studies, and the statistical tools used in the analysis.

An important working assumption for the present study is an understanding of grammar as a primarily social entity, very much like de Saussure (1983 [1916]: ch. III, § 2) had formulated it: "[The language system] is the social part of language, external to the individual who by himself is powerless either to create it or to modify it. It exists only in virtue of a kind of contract agreed between the members of a community." This position is close to what Labov (2010: 7) describes as the *central dogma* of sociolinguistics, namely, the priority of the community over the individual in linguistic analysis.[1] The mere existence of grammatical taboos can only be explained with reference to the heterogeneity of communicative situations (register differences within speakers) and, most of all, the hierarchical stratification of society, how it is reflected in language, and how speakers, diverse as they are, deal with it. De Saussure's metaphor for the language system as a kind of implicit *contract* between the members of the community is more helpful for an understanding of these issues than a purely cognitive view on grammar. Grammatical taboos, paradoxical as they are, are a part of this "contract".

Another motivation for my emphasis on the social nature of grammar is methodological. The statistical tools used in experimental morphosyntax treat the data gathered in an experiment as *random samples* from a larger *population*. In our case that population is the language community.

All inferences drawn this way are therefore necessarily inferences about the community, if anything. Thus, in experimental morphosyntax we are establishing facts that are primarily *social* facts. They are, secondarily, psychological facts insofar as any social facts have their base in the individuals, their attitudes and behaviour. But the acquisition and practise of language by each individual speaker is mediated by his position within society. There is, therefore, no reliable way to infer from a single speaker to the totality of the language *by empirical means*.

1 In a similar vein, Devitt (2006) argues that linguistics is not, and cannot be, psychology, but studies "linguistic reality". I agree, if the latter is understood as a social reality.

2.1 Empirical grammaticality

The increasing use of experimental research methods in grammatical theory over the last twenty years has called into question a stance towards linguistic data which by and large is based on expert knowledge and expert consent. But the enthusiasm of the early revolutionary phase[2] has somewhat cooled down recently, due mainly to the insight that informal expert judgements and results from grammaticality experiments on the same data usually converge, as has convincingly been shown by Sprouse et al. (2013) who report a convergence rate of about 95 % for a huge sample of grammaticality judgements from ten volumes of *Linguistic Inquiry*.

I have no doubt that most expert judgements that can be found in the literature are sufficiently reliable. Nevertheless, the study by Sprouse et al. (2013) shows less than what a naïve reader might think. This has to do with limitations of the statistical tools that are usually, and still, applied in empirical linguistic studies.

Experimental morphosyntax is a descendant from psycholinguistics and psychology. It therefore inherited the shortcomings and limitations of the research methods used in those disciplines. The culture of "Null hypothesis Significance Testing" (NHST, Cohen 1994) that has dominated these fields for more than fifty years has always received some criticism.[3] It only recently started to lose its dominance in the behavioural sciences[4] – with linguistics being a bit behind. The problem with NHST is, first of all, that null hypothesis testing was usually practised as the *only* way of doing inferential statistics in the behavioural sciences and linguistics, and it was usually done in a particularly trivial way.

Null hypothesis testing *enforces* an operationalisation of research hypotheses in a way that allows for testing the "nil hypothesis" (Cohen 1994; Kline 2013) that there is *no contrast between two data samples*. This is all too often the only hypothesis that really is being tested: the famous "p-value", the main result of null hypothesis testing, is the probability of the data under the assumption that the null (i. e., "nil") hypothesis is true. The standard in the behavioural sciences is that *significance* is reached with $p < .05$. The null hypothesis can then be rejected.

2 Schütze (1996) and Cowart (1997) may count as the pioneering works on experimental morphosyntax. More recent conceptual work in the field focuses on core methodological questions like the equivalence of different experimental methods, issues of gradient acceptability and the relation of theoretical and experimental morphosyntax. A non-exhaustive sample of this important work includes Featherston (2005, 2007, 2009), Weskott and Fanselow (2009, 2011), Bader and Häussler (2010), Sprouse (2011), Sprouse et al. (2013), Sprouse and Almeida (2017).

3 See, besides Cohen (1994), for instance, the criticism of the "null ritual" by Gigerenzer et al. (2004).

4 See Kline (2013) for a summary of the debate, its history and its main arguments, and Vasishth et al. (2018) for a recent contribution from psycholinguistics in that spirit.

Several conceptual problems arise here that are rarely discussed in linguistics. First, the nil hypothesis is always false anyway: it is absolutely unlikely that two different sentences have *exactly* the same corpus frequency or *exactly* the same probability of being accepted. If there is a contrast, however, it is only a matter of *sample size*, whether the p-value is below .05. For very small contrasts, an experiment might need many thousands of participants, but, with sample size increasing arbitrarily, it will always be possible to get a significant contrast.

What empirical linguists therefore should talk about, but rarely do, is the – predicted and measured – *size* of contrasts. In his pioneering work on *effect sizes*, Cohen (1988) introduced various standardised measures of effect size for different types of data. He also proposed rules of thumb for the classification of effect sizes as "small", "medium", and "large" effects. Especially relevant for our discussion is his characterisation of a medium effect size, of which he states that it is "likely to be visible to the naked eye of a careful observer" (Cohen 1992: 156).

Sprouse and Almeida (2017) show that the great majority of the contrasts that they observed fall into the category of medium or large effect sizes. An expert, certainly counting as a "careful observer" in Cohen's sense, usually does not depend on experiments to establish linguistic facts – her "naked eye" is sufficient. Many of the observed effects even belong to the category of *large* effects and should be hard to overlook to anybody. This explains the high correlation of expert judgement and experiment participants in the study by Sprouse et al. (2013). The contrasts mostly discussed in grammar books and morphosyntactic theory are large enough to be beyond doubt even to the ordinary language user.

The expert linguist classifies sentences not only into the categories "grammatical" and "ungrammatical", but also identifies coarse-grained levels of degraded grammaticality. Usually, two levels of weaker and stronger markedness are identified, symbolised by "?" and "??". The important point here is that linguists tend to agree on these judgements. This indicates that these contrasts are more than subjective impressions and presumably have at least medium effect size. This leads us to our first *hypothesis on relative acceptability* in (1).[5]

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

H I will both be put to test in the study introduced below and used for the categorisation of grammatical taboo phenomena.

5 Where necessary, I am using the symbol '✓' for "perfect", unmarked grammaticality.

As already mentioned, contrasts in acceptability between structures, often minimal pairs, are what is tested in NHST in linguistics. Some authors (like Sprouse et al. 2013) even consider this as the most adequate way of testing linguistic hypotheses. I disagree. Grammaticality is an absolute, not a relative property of a sentence. Speakers judge the acceptability of a sentence without recourse to other sentences. Nobody would agree to a statement like "Sentence B is grammatical, if compared to sentence A, but ungrammatical if compared to sentence C".

What we should do, and can easily do, is deriving hypotheses for *absolute* acceptability values: an expert expects participants of an elicitation experiment to share her intuitions. Thus, if S_u is ungrammatical according to the expert, subjects should rate it as ungrammatical, too: S_u should be judged acceptable in 0 % of the trials. Likewise, a perfect, unmarked sentence S_p should be judged acceptable in 100 % of the trials.

Now, we do know that such extreme expectations are unrealistic. Subjects in experiments accidentally make mistakes and the design of an experiment itself might induce artefacts. Hence, we should liberalise these expectations a bit. I am using 10 % here: "unmarked" then translates into acceptability > 90 % and "ungrammatical" into acceptability < 10 %. In addition, we should add a zone of uncertainty of, say, another 10 %, such that a hypothesis can be considered as severely challenged if acceptability is below 80 % for S_p or above 20 % for S_u. That is: in those cases the expert who wants to uphold her claims must provide a convincing explanation for the results.

The range between 20 % and 80 % straightforwardly is the domain of markedness, where slight markedness is in the upper, and more severe markedness in the lower half of the spectrum. I again put a zone of uncertainty in between. I chose to use a broader range for '??' than for '?' under the assumption that higher markedness produces more variance than lower markedness which is counterbalanced by a wider corridor. These considerations are summarised in the general hypothesis about absolute acceptability (2).

(2) **H II: 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

The hypotheses H I and H II perhaps formulate the maximum of the empirical hypotheses that one can derive from a morphosyntactic analysis developed in the linguist's "armchair". What is usually put to test, however, is only a small fraction of what would be necessary, the weakest of all possible claims: there is some difference – however small it actually is – in the proposed direction between two test conditions. But such an utmost minimum of statistical inference is not providing evidence for or against morphosyntactic analyses.

H II will also be put to test in the study introduced below, and it will be used for the categorisation of grammatical taboo phenomena. The ranges proposed for the different categories of gradient grammaticality should be understood as a starting point for further exploration and critical examination. Implicit in the above considerations is an idea of *empirical grammaticality*. I define it as in (3).

(3) **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 here understood as the probability by which an expression is judged acceptable. By assumption, every imaginable expression has a specific grammaticality value in the speech community. It is being estimated in morphosyntactic experiments. The categorisations of grammaticality used in linguistic theory (like √,?,??,*) are *ordinal* in nature. But the measures that are being used in empirical linguistics are often continuous.[6] In (1) and (2), I made a proposal about the relation between these two different scale types. Empirical grammaticality is at the heart of grammar research. It is the subject matter both of grammar modelling and of morphosyntactic experimentation.

What perhaps everyone (even those who are critical of empirical linguistics) can agree on is that experimental methods are useful where they *complement* expert judgement, especially for research questions where expert judgements are *insufficient*. This can have *sociological* or *psychological* reasons.

Sociological motivations arise whenever a syntactic construction is controversial within the community, or at least among the experts. In such a case, an expert

6 Rating scales have a pre-defined minimum. If they are given a numeric interpretation, they are ratio-scales, which implies, among other things that statemens like "the grammaticality of S_1 is twice as high as that of S_2" are meaningful. This describes current practice in experimental linguistics, as I understand it. I am a bit sceptical whether this is a reasonable way of operationalising the intuitions of linguists and speakers, which presumably are beyond doubt only at the ordinal level.

judgement is just one *opinion*. Insistence on the privileged (expert) status of some linguists' opinions within the community would be tantamount to a *prescriptivist* position (which modern linguists reject). Linguists who rely on expert judgements run the risk of inadvertently upholding a prescriptivist position, when they fail to recognise the sociocultural factors behind (their own) acceptability judgements.

We have a psychological motivation for an experiment whenever a judgement is too subtle for a definitive expert judgement, e. g. in the case of gradient acceptability with small contrasts, or when the precise level of acceptability becomes relevant. Reliance on expert opinion limits the empirical range of phenomena to medium to large effects that are visible to the naked eye of the observer. There is no reason to assume that all grammatical phenomena produce such effect sizes. The scale of gradient acceptability that the expert is able to establish without experimentation might be too coarse-grained to cover all relevant distinctions.[7]

The study that I am presenting here has both kinds of motivations: *Grammatical taboos* are cases of *socioculturally induced markedness*, i. e. cases of reduced, gradient grammaticality with a sociocultural origin.

2.2 Measures of empirical grammaticality

The task that is most often used, and to my mind also most adequate, in the designs of morphosyntactic experiments, is the acceptability judgement where subjects are presented expressions, mostly sentences, in isolation and asked to rate them. The important advantage of elicitation over production oriented methods lies in the fact that grammar researchers are interested not only in what is actually realised, but what could be realised, the potential of a language system.

There is a sociolinguistic dimension, insofar as subjects may judge expressions as acceptable which they might never use themselves, i. e. these experiments also measure the subjects' level of tolerance towards the language of others. The set of expressions a speaker accepts should be a superset of those she uses herself.

Experiments may differ in details like presentation mode (written or auditory stimulus presentation), offline (questionnaire) or online (speeded acceptability judgement) task, and character of rating scale (non-gradient binary "yes-no" scale vs 4-7 point rating scales). A very important recent finding is that experiments using these different designs broadly converge in their results (Bader and Häussler 2010; Weskott and Fanselow 2011). They may therefore serve equally well for the estimation of the empirical grammaticality of an expression.

7 This implies that the scale used in hypotheses I (1) and II (2) should be subjected to critical examination and, if necessary, revision.

The experiment presented in Section 4 is an offline written questionnaire task with a 7-point rating scale where only the endpoints of the scale are labelled, as illustrated in (4). It is common practice to analyse the ratings in such experiments not only ordinally, but numerically, whereby it is assumed that the distances between the points of the scale are the same. In the study presented here, the 7-point scale is mapped onto the probability scale, as shown in (4).

(4) *Values for the 7-point scale used in this study*

highly possible □ □ □ □ □ □ □ impossible

$1 \quad \frac{5}{6} \quad \frac{4}{6} \quad \frac{3}{6} \quad \frac{2}{6} \quad \frac{1}{6} \quad 0$

This allows us to calculate means, variances, standard deviations and other continuous measures for the tested items. The means with their confidence intervals can be used as measures of *absolute acceptability* to test hypothesis H II.

For testing hypothesis H I, we need to calculate the effect sizes of contrasts between two test items. The best-known standardised measure of effect size for continuous data is *Cohen's* d. It is calculated as the difference between the means of two samples, divided by their pooled standard deviation (SD).

We have a small effect with $d > .2$, a medium effect with $d > .5$ and a large effect with $d > .8$. As Cohen (1988, 1992) noted repeatedly, these classifications should be used with care, as sort of rules of thumb. Every discipline should come up with its own conventions for the interpretation of contrasts. In the absence of other proposals, I will stick to the categories Cohen introduced. At least, they are a reasonable starting point.

An important prerequisite for the use of Cohen's *d* is that the two populations from which the samples stem can be considered to have the same variances. Only under this condition is pooling of the two SD's reasonable. We will see below that this condition cannot be met with grammatical taboo phenomena which seem to have as one characteristic that their variance differs from that of other kinds of markedness. Therefore, I am not using Cohen's *d* to test proposals related to H I.

Instead, I will rely on the ordinal character of the rating scale and use a measure of effect size for ordinal data, *Cliff's delta* (Cliff 1996). It 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.[8]

8 For its calculation with independent data sets all pairs (x_i, x_j) 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_j, or whether they are equal. Cliff's delta is the difference between the number of pairs where x_i is higher and the number of pairs where x_j is higher, divided by the total number of pairs.

Romano et al. (2006) compared Cliff's delta with Cohen's *d* and made a widely accepted proposal for small, medium and large effect sizes that I will use in the discussion below: $d > .147$ is a small effect size, $d > .333$ a medium effect size and $d > .474$ is a large effect size. These rules of thumb might be in need of revision to match the empirical situation in experimental morphosyntax.

2.3 Further goals

Most work in experimental morphosyntax is surprisingly neutral with respect to the problem that the target language usually is the *standard language*. Especially in the German language community, the standard language has its own history as the prestige variant used primarily in written language. Linguists, on the contrary, are interested in unbiased judgements related to speakers' native variants, i. e. informal spoken language. But the participants of experiments usually have 10+ years of school training in the standard language behind them, after which it would be quite naïve to expect judgements based on pure native speaker intuition.

This is where the present study tries to shed some light on. The focus lies on two particular research questions:

a) Is it possible to minimise or neutralise the influence of linguistic ideologies in our data?
   For anybody working in experimental morphosyntax, it is important to *understand* the empirical effects of ideologies in our experiments. But perhaps it might also be necessary (for everyone doing grammar research) to *accept* them as natural part of grammars as sociocultural entities.
b) Can we distinguish by empirical methods between ordinary grammatical markedness (with grammar-*internal* cause) and the markedness of grammatical taboos (caused *externally*)?
   As we will see below, this question is likely to receive a positive answer: A mean acceptability value in the range of "??" ('marked') could come about under greater or smaller uniformity among participants. The former seems to be more typical of internally caused markedness, the latter of grammatical taboos.

Another goal of this study is the exploration of the feasibility of the general hypotheses I and II posited in (1) and (2) and the concept of empirical grammaticality.

Under the convention that in calculating the numerator of *d* the smaller value is always subtracted from the larger one, *d* takes on a value between 0 (no difference in probabilities) and 1 (maximum difference in probabilities).

The answer will again be rather positive. Before we can go into the details of the experiment, I will introduce grammatical taboos and discuss their role in the German standard language in the following section.

3 Grammatical taboos in standard German

Standard German (SG) is the *prestige variety* of the German language. This is a result of its historical origin as the written (and for a long time only written) variety. This prestige status is *at odds* with the quite recent development of SG as preferred variety in informal oral language use.

I assume that the requirement to keep distance between written and spoken language is one core principle of the *German standard language ideology*:

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

While the GSLT had an important sociocultural function in establishing a supraregionally comprehensible written variety, it has today lost its motivation due to the decline of the traditional regional dialects. Today's reality is that SG is also used in spoken informal communication: in an informal register which nevertheless hugely overlaps with the formal register of SG.

It seems, though, that the common view of SG in the society, which also dominates in the educational systems, still implies a sharper division between formal and informal, written and spoken language, insisting on adherence to the GSLT. Given the huge overlap of the variants, this seems impossible.

In practice, adherence to the GSLT[9] 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:[10]

9 Which, in fact, means that the linguistic community is struggling to retain the *illusion* of fulfilling the GSLT.

10 While I find the notion "grammatical taboo" quite appropriate, I am not the first one to use it. The earliest use that I could find is by McBryde (1943). In his reconstruction of the history of SG, Weiß (2004) uses the notion *artefact* for both types of shibboleths. His focus lies in an evaluation of the unnaturalness or implausibility of these artefacts, ranging from rather harmless taboos over natural options to typologically highly improbable rules of the SG grammar, and even rules that contradict language universals.

(6) **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.

This study focuses on grammatical taboos. Speakers usually do not evaluate different registers of a language as equal, 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 (7).

(7) **Paradox of grammatical taboos**
a. 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.
b. Because of the taboo over C, speakers of L who conform to the taboo nevertheless *believe* that C should not and therefore does not belong to L.

From the axiom that grammars are sociocultural entities, it follows that they develop and change over time due to the linguistic practice of the community. Adherence to a grammatical taboo by a great majority of the community, as one aspect of language use, can therefore have an influence on the historical development of the language system. Taboos are, thus, a vehicle for language ideologies (ideologues) to shape language according to their "will".

Standard German has a special history and function within its community, as compared, in particular, to Standard English. As Durrell (1999) describes, the standard language notion as it is usually understood in the German speaking societies (including German (socio-)linguistics) restricts it to the language of social groups with higher education and/or higher socioeconomic status. The exclusion of everyday informal language (of less educated social groups) from the standard language that follows from this is not paralleled in the English speaking societies.

The English standard language developed on the basis of a particular spoken variant practised in the political centre. Therefore, there was a *parallel development* of written (formal) and spoken (informal) registers of the standard language from the 16th century on. The German standard language, on the other hand, existed

for a long time *only in the written, formal register*. Dialects prevailed in spoken communication until the middle of the 20th century in Germany. We thus envisaged a situation of diglossia, as we still do nowadays in the German speaking parts of Switzerland.[11]

Since the middle of the 20th century, an informal register of standard German has developed that today can be seen as the native language of the majority of speakers in Germany.[12] Reference grammars of German spend much effort on describing the German written standard in many details. A systematic reference grammar that focuses on the grammatical specifics and variants of this informal register of SG is still to be written.[13]

Von Polenz (2000, 1994, 1999) in his outstanding three-volume work described the development of SG (aka New High German, NHG) at length. For a long time, SG has been the project of a small elite, the minority of literate people who in the beginning of the 16th century made up less than 5 % of the population (von Polenz 2000: 128).[14] Weiß (2004) shows that this situation led to the introduction of *artefacts* into the NHG grammar by prescription, some of which are still present. Examples that Weiß gives are a. o.: the elimination of negative concord; the stigmatisation of analytical inflection due to an ideological position which ascribes a primordial and therefore preferable status to synthetic inflection; the conservation of outdated forms of synthetic inflection for the same reason.[15]

An especially intriguing case is the stigmatisation of the auxiliary use of the verb *tun* 'do' which has been explored in great detail by Langer (2001) and Davies and Langer (2006). Based on analyses of grammars from the relevant time peri-

11 Durrell (1999) also claims that the German standard language is codified to a higher degree than is the case for English. This claim is disputed by Davies and Langer (2006: 269–276).

12 Weiß (2004, 2005) describes this development in some detail and points out its relevance for grammatical theory.

13 Several projects are underway that come close to this task. One is "Korpusgrammatik – grammatische Variation im standardsprachlichen und standardnahen Deutsch" by the Institut für deutsche Sprache (IDS): http://www1.ids-mannheim.de/gra/projekte/korpusgrammatik.html, another is the "Atlas zur deutschen Alltagssprache" 'atlas of informal German language'), carried out by Stephan Elspaß, University of Salzburg and Robert Möller, University of Liège: http://atlas-alltagssprache.de.

14 There also was an urban-rural discrepancy, with literates up to 10 % in towns (von Polenz 2000: 128).

15 The prescriptive literature of those days belongs to what Milroy and Milroy (1985) in their classical monograph called the *complaint tradition* in the development of standard languages. While their focus is on the history of Standard English, complaint traditions occur in many, if not all cultures that develop a linguistic standard. They seem to be an unavoidable by-product of linguistic standardisation, at least in the hierarchical societies of our times with elites that feel a need for sociolinguistic distinction.

ods, they reconstructed the history of this stigmatisation as proceeding in five phases:

1. No stigmatisation of auxiliary *tun* can be found in the grammars until 1640;
2. Poetic grammars stigmatise the construction for poetic texts (1640–1680);
3. The stigmatisation spreads further from here. Between 1680 and 1740, grammars mention the construction as bad in formal written language;
4. From about 1740 on, grammars stigmatise the construction as an aspect of lower class speech, and as not belonging to the written standard;
5. At the beginning of the 20th century, it is often recognised in grammars that auxiliary *tun* may serve some purpose, especially in cases of verb fronting in the simple tenses. In this situation, the fronted verb has to be infinite and another finite verb is needed to fill the verb-second position:

(8) *Wissen tut sie das nicht.*
know-INF do-3SGPRES she it not.

The Duden grammar (Duden 2016: 435) allows for auxiliary *tun* exactly and only for this case. Davies and Langer (2006) show that the issue is still controversial. Some style guides even show reservations about structures like (8), and the same can be said to some extent about school teachers whose attitudes towards a number of stigmatised constructions have been explored by the authors.

The historical development described in the studies on auxiliary *tun* by Langer (2001) and Davies and Langer (2006) is a major inspiration for this study. It suggests, as indicated in (9), an implicational relation between different types of acceptability judgements: an aesthetic judgement is the most rigorous, followed by judgements based on general norms for written language. In comparison to these two judgement types, the linguists' perspective is the most liberal: is this a possible expression of your language, considering all registers of the language, in particular the informal spoken language of everyday usage?

(9) beautiful (poetic) language ⊂ norm-compliant language ⊂ informal language

This implication also suggests that beautiful language is always norm-compliant and norm-compliant language always useful in everyday conversation. While this certainly is not true, (9) can be understood as another ingredient of the German standard language ideology, i. e. speakers might assume it to hold and base their acceptability judgements on it.[16] In the study presented below, the scale in (9) is

16 One effect of this has been a strong bias towards the prestige variety in the tradition of teaching German as a foreign language, that lasted for a long time.

used to compare different types of acceptability judgements and their effects on the rating of grammatical taboos.

Is it possible to uncover with empirical means the paradoxical nature of grammatical taboos? This is one of the questions that underlie the present study. Is it possible to adjust our elicitation methods – using the scale in (9) – such that one or the other side of the paradox described in (7) is neutralised? Grammatical taboos, then, could be identified by this property, whereas rules that are not taboos should be rather neutral in this respect.

The intensity of the stigmatisation of grammatical taboos differs from case to case. In this study, I am comparing two presumably strong taboos with two rather weak ones. The case of auxiliary *tun* has already been introduced (10-a). It surely is one of the strongest grammatical taboos in German. The second strong taboo phenomenon explored here are German sentences introduced with the connector *weil* 'because' with second position placement of the finite verb ('weil V2-clause', 10b). Subordinate clauses usually have clause-final verb placement. This also holds of subordinate clauses with *weil*. However, there is a non-standard use with V2-clauses which occurs predominantly in the spoken informal register. It can be traced back at least to the 19th century (Elspaß 2005; Elspaß 2010).[17]

(10) 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

In an empirical study with German teachers and linguists, Davies and Langer (2006: 148–154) tested the recognition of 14 more or less stigmatised phenomena of SG. The two phenomena in (10) were the ones with the highest score in this study. I take this as evidence for the hypothesis that these are strong taboos.

They will be compared with two weaker ones. I assume them to be subject to the general standard language taboo, as they are found primarily in spoken informal conversation. On the other hand, I assume that ordinary speakers do not try to avoid these particular constructions as consciously as they presumably do

17 Selting (1999) shows that earlier versions of the *weil* + V2-construction are attested from the Old High German to the end of the Early New High German phase, but it seemed to have vanished in New High German until its reoccurrence in the 20th century. She assumes a historical continuity from these earlier constructions to today's use of *weil* + V2, whereby the construction survived in informal oral communication. The evidence by Elspaß is in line with this hypothesis, but there still is a gap of about 200 years, between the 17th and the 19th century, where the construction is rarely, if at all, attested.

with auxiliary *tun*. They were not included in the sample of phenomena tested by Davies and Langer (2006). Apart from that, the classification of the two phenomena as weak taboos is a research hypothesis to be tested in the presented study.

The first weak taboo is the double perfect construction which consists in the recursive application of the perfect construction. A simple perfect contains an auxiliary (*sein* 'be', or *haben* 'have') and a perfect participle of a full verb (11-a). The double perfect is based on a simple perfect with the auxiliary itself put into the perfect construction, such that we have the auxiliary occurring twice, as finite verb (or infinitive) and as perfect participle, plus the participle of a full verb.

The 8th edition of the Duden grammar (Duden 2009: 514) judges the construction as incorrect for the written standard language and reserved for informal language. This restriction has been eliminated in the 9th edition (Duden 2016: 526), though it is still emphasised that the construction is used primarily in informal registers. There is a substantial amount of literature about it, so that its history and function are quite well understood (see Zybatow and Weskott 2018 for a summary). The construction is attested in corpora of written German, though with very low frequency. The double perfect shares with the auxiliary *tun* construction that it belongs to the German language for a couple of centuries already. Contrary to the latter, the double perfect has enjoyed less attention in the public discourse, although all "logical" arguments against auxiliary *tun* could also apply here: it can be seen as basically a "redundant" alternative to the past perfect construction, to be considered less "elegant" due to the double occurrence of the auxiliary.

The other weak taboo is the pronominal use of so-called *d*-pronouns (11-b). Barbour and Stevenson (1998: 147) use this phenomenon to exemplify the strong bias towards formal written language in German grammars, because *d*-pronouns are barely ever mentioned, although they are regularly found in the speech of even educated speakers.[18] The recent years saw a certain amount of substantial work on this phenomenon (see Portele and Bader 2016 for a summary). Pronominal uses of the *d*-pronoun, even those without demonstrative force as in (11-b), are summarised as "demonstrative use" in the Duden grammar (Duden 2016: 280f) and by most authors.

Bosch et al. (2007) report in their corpus study a huge difference between written language, where *d*-pronouns are rather infrequent, and spoken language, where they occur much more often, especially when they are sentence-initial. The context in which *d*-pronouns occur in our experiment (11-b) is predicted to be a context where *d*-pronouns are rather dispreferred. Markedness of this construction

18 The authors mention the early exception of Eisenberg (1986: 191) for a special demonstrative use. A first substantial discussion can be found in Zifonun et al. (1997: 558–559).

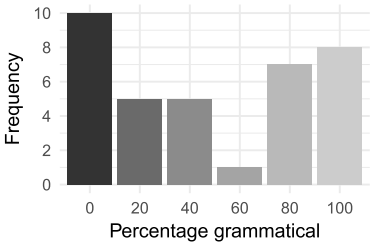

Figure 1: Distribution of experiment participants in judging sentences like (12) (5 items) in the experiment by Bader and Schmid (2006).

might therefore not only be due to the GSLT, but also due to the grammar of *d*-pronouns.[19]

(11) Non-salient grammatical taboos (explored in the experiment)

a. *Als Peter kam, hat Max bereits gewonnen gehabt.* dbl. perfect
When P. came has M. already won had

b. *Als Paul kam, neckte ich den.* *d*-pronoun
When P. came teased I him

Some of these phenomena have already been subject to empirical investigation, though with special experimental designs and different objectives. In particular, the dimension of the stigmatisation of these constructions, which is the focus of the present study, is not explored systematically.

Bader and Schmid (2006) studied several uses of auxiliary *tun* in a speeded grammaticality judgement task (binary yes/no decisions) with visual word-by-word presentation. A condition with object fronting, as in (12), led to a non-normal distribution of the judgements as in Figure 1 (after Bader and Schmid 2006, figure 1).

(12) *Den teuren Schmuck tut Monika sicherlich verstecken.*
the expensive jewelry does M. surely hide

Bader and Schmid (2006) interpret this as a bimodal distribution that gave rise to splitting the participants into two groups with high (89 %) and low (15 %) mean acceptability ratings for sentences like (12).

The high variance and disagreement between subjects that is observed here might be a characteristic effect of grammatical taboos: external (ideological) factors are less consistently shared among speakers and much more controversial than truly grammatical factors which are internal to the language system. The

19 I do, however, suspect that grammatical analyses of *d*-pronouns are still somewhat biased as they are based primarily on written language use. Structures like (11-b) have not yet been investigated experimentally.

division among speakers could also result from the paradox of grammatical taboos: when participants make a judgement, they have to decide which side of the taboo they prioritise and may make different choices, although they share the same (paradoxical) opinion about the construction.

A few studies on *weil* V2-clauses also suggest for this construction the status of a marked construction (Volodina 2009; Antomo and Steinbach 2010). In the same direction points the study by Zybatow and Weskott (2018) on the double perfect. Empirical work on *d*-pronouns focuses on the different resolution strategies for personal and *d*-pronouns, irrespective of issues of acceptability (see, a. o., Bosch et al. 2007; Bosch and Umbach 2007; Portele and Bader 2016).

There is also at least one neurolinguistic study on grammatical taboos: Hubers et al. (2016) have been able to detect neurophysiological correlates of the paradoxical nature of linguistic stigmatisation in language comprehension. Speakers with a "puristic" attitude towards language show fMRI patterns in parsing norm violations which are typical of grammatical sentences, but in addition show a component that is usually found with repair/correction phenomena.

4 The experiment

The present study has two main objectives: the exploration of the empirical effects of grammatical taboos, and an evaluation of the concept of empirical grammaticality as introduced in Section 2, including the hypotheses H I and H II. The objectives are connected in that the clarification of empirical grammaticality makes it easier to understand the empirical effects of grammatical taboos.

The experiment is an acceptability rating study with a written questionnaire. Judgements were given on a 7-point rating scale where only the extremes are labelled. The subjects for the experiment were randomly selected from the participants of introductory courses in German studies at the University of Bielefeld. Students were rewarded with course credits for their participation. Here again, I followed typical practice in experimental morphosyntax. With respect to the tested grammatical taboos, there are four empirical hypotheses:

(13) **H III: Hypotheses on the acceptability of grammatical taboos**

a. Grammatical taboos are marked, not ungrammatical.
b. They differ gradually from each other, depending on their salience.
c. The acceptability of grammatical taboos is dependent on instruction type to a higher degree than judgement of ordinary morphosyntactic markedness.

d. Grammatical taboos differ from grammar-internal markedness in that they show greater between-subject variance.

The chosen design, especially the bias towards formal standard language due to the use of a written questionnaire, might be working against these hypotheses, but if they can be confirmed even under such disadvantageous conditions, the argument in their favour is even stronger.

The study presented here is only a starting point for the research program sketched in the previous sections, and in most aspects has to be seen as explorative.

4.1 Materials

The test sentences were distributed in random order on a written questionnaire. The experiment was divided into three subexperiments that use exactly the same stimulus material, but differ in the instruction. Each participant took part in only one of the three subexperiments.

The three subexperiments were using an aesthetic judgement, a norm-oriented judgement and a possibility judgement, respectively. On top of the first page of the questionnaire was an instruction text that described the intended judgement type a bit closer (see the first section in the appendix that is available for the online version of this article). Every sentence was followed by a judgement scale where the left and right ends were labelled, according to the judgement type of the subexperiment, as in (14).

(14) a. Aesthetic judgement (type "A"):
sehr schönes Deutsch □ □ □ □ □ □ □ *kein schönes Deutsch*
very nice German no nice German

b. Normative judgement (type "N"):
sehr gutes Deutsch □ □ □ □ □ □ □ *kein gutes Deutsch*
very good German no good German

c. Possibility (linguistic) judgement (type "P"):
sehr gut möglich □ □ □ □ □ □ □ *unmöglich*
highly possible impossible

The repetition of the labels after each item is crucial to remind subjects of the intended judgement type. We constructed 8 lexical variants of the four taboo phenomena, using identical structures but different content words (verbs, nouns, adverbials etc.), in a 2x2 design (for the double perfect it was 3x2, see below). Each taboo sentence was paired with a semantically equivalent variant that avoided the

taboo violation. For both of these sentences we constructed clearly ungrammatical variants where ungrammaticality was due to an agreement error on the finite verb. Examples for the ungrammatical conditions are here only displayed for auxiliary *tun*.

(15) **Strong taboos**[20]

a. Auxiliary *tun*'do'[21]

(i) *Damals **tat** Hans gut lesen.* +gramm,–taboo
Then do-PST3SG H. well read

(ii) *Damals hat Hans gut gelesen.* +gramm,+taboo
Then have-3SG H. well read

(iii) *Damals tat**en** Hans gut lesen.* –gramm,–taboo
do-PST3**PL**

(iv) *Damals hab**en** Hans gut gelesen.* –gramm,+taboo
have-3**PL**

b. *Weil* 'because' V2-clauses
Tim humpelt , ...
T. hobbles

(i) ***weil** er **hat** seinen Fuß gebrochen.* +gramm,–taboo
because he has his foot broken

(ii) ***weil** er seinen Fuß gebrochen **hat**.* +gramm,+taboo
because he his foot broken has

(16) **Weak taboos**

a. *d*-pronouns
*Während Felix sprach, schaute ich **den**/ihn an.*
While F. spoke looked I him at

+gramm,–taboo = *den* / +gramm,+taboo = *ihn*

20 In the coding of the experiment conditions, "+" means constraint fulfilment and "–" means constraint violation, so that e. g. a "–gramm/–taboo" condition violates a grammatical rule and contains a taboo item.

21 Auxiliary *tun* is used in both simple tenses, simple past and simple present. The simple present use is presumably more widespread and therefore perhaps the better candidate for such an experiment. Our choice for the past tense is due to the decision to construe minimal pairs that are both syntactically and semantically equivalent, as much as possible. The present perfect (15-aii) is regularly used in a past tense interpretation and it is an analytic tense, therefore the ideal candidate for an equivalent non-stigmatised expression. For the present tense use of auxiliary *tun*, no such neutral analytic alternative is available.

b. Double perfect

(i) +gramm/–taboo, version 1:
Als Peter kam, ***hat*** *Max bereits* ***gewonnen gehabt***.
When P. came has M. already won had

(ii) +gramm/–taboo, version 2:
Als Peter kam, ***hatte*** *Max bereits* ***gewonnen gehabt***.
When P. came had M. already won had

(iii) +gramm/+taboo:
Als Peter kam, ***hatte*** *Max bereits* ***gewonnen*** .
When P. came had M. already won

The double perfect construction exists in two versions, with the finite auxiliary in present tense (16-bi) and past tense (*double past perfect* [16bii]). Both variants are tested in the experiment.

Eight lexical variants were constructed for each of the four taboo phenomena with 32 (= 4 x 8) test sentences for auxiliary *tun*, *weil*-V2 and *d*-pronouns and 48 (= 6 x 8) test sentences for the double perfect. Each of the eight lexical variants occurred only once per questionnaire, with two items per condition for auxiliary *tun*, *weil*-V2 clause and *d*-pronoun. Only two of the six conditions for the double perfect were presented with two items per condition on a questionnaire, but this was altered systematically so that every condition and all items were presented to the same amount. The acceptability of the two versions of the double perfect differed only minimally, however. This allowed us to combine the two versions in our comparative analyses.

Included in the test materials were 64 test items of an independent experiment on verbal complexes, 16 per questionnaire, and 48 filler sentences. Among the filler sentences were 21 ungrammatical sentences, nine sentences which are syntactically marked according to standard criteria,[22] and 18 sentences that contained so-called anglicisms and were otherwise morphosyntactically unmarked. These sentences with anglicisms are expected to be judged as degraded only by subjects with reservations towards loan vocabulary from English. Stigmatisation of such forms is frequent in public reflections on language.

Every questionnaire contained 96 sentences. The test material was distributed over 12 different variants of the questionnaire. The sentences were presented in

22 Section 3 in the appendix, available only in the online version, describes the material used as marked fillers. Two of the marked sentences were identical. They were included to control for consistency of judgements. The sets of ungrammatical and marked sentences comprise a variety of different phenomena that are described as marked or ungrammatical in the linguistics literature.

pseudo-randomised order which varied systematically over the 12 versions to avoid effects of order. The questionnaires were used four times in each judgement type, so that all in all 144 questionnaires were filled in, 48 in each judgement type.

4.2 Statistical methods

The elicited judgements were given a numerical value on the scale between 0 and 1 in order to match the probability scale that is used in our definition of empirical grammaticality (17). For the sake of readability, I will nevertheless use labels from the integer scale from –3 to 3 to refer to particular levels of the scale in tabular presentations and diagrams.

(17) *Labels and values for the 7-point scale in the following discussion*

label:	*–3*	*–2*	*–1*	*0*	*1*	*2*	*3*
	□	□	□	□	□	□	□
value:	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

Inspection of the data does not justify the assumption of huge inter-individual differences in their use of the 7-point scale for about 90 % of the participants. But 14 of the 144 participants showed a deviating behavior of avoiding the extremes (the so-called central tendency bias). Their judgements have been excluded from the data.[23]

Missing responses were also removed. In addition, I decided to remove obvious errors from the data. Those obvious errors were cases where uncontroversially ungrammatical sentences were judged with the highest rating "3", and conversely, unmarked perfectly grammatical sentences with the lowest rating "–3".[24] The overall theoretical maximum of observations of 144*96 = 13 824 was thus reduced to 12 353 observations (89 %) that could actually be analysed.

For the statistical analysis of contrasts, estimation of effect sizes etc., a complication has arisen. An important precondition for the application of the usual analytical tools is not met: *homogeneity* of the variances of the compared data

23 Those subjects avoided the extreme rating values like "–3" and "3", but also with a bias such that low ratings have been avoided more consistently. The 14 participants have in common that they rated the ungrammatical filler sentences with a mean higher than 0.33. While leaving their data in would only add more noise to the data without removing the observed effects, they were distributed unevenly among the three judgement types.

24 This only had the effect of reducing noise in the data. The thus excluded data points were about 0.5 % of the observations.

sets. *Change in variance patterns among subjects* obviously is part of the effects of prescription. Both the prescriptively oriented judgement types and stigmatised constructions provoke a kind of heterogeneity among participants that is different from the variance that occurs with "natural" judgement and sentence types. Section 4.4 summarises the crucial observations in this respect in a comparison of grammatical taboos with ordinary markedness. Variance heterogeneity becomes visible especially when the results are aggregated subject-wise, whereas the effect remains somewhat hidden, as long as the raw judgement data are inspected. The methodological consequences of this result are discussed below.

The participants were first-year students of German studies. There was a huge majority of female participants which is normal in German studies courses (and in many linguistic experiments). Effects of gender differences which might be relevant here could therefore not be investigated systematically.[25]

The group of second language speakers among the participants was small, but their judgements did not differ greatly from the others, so there was no reason to exclude them (see Table 1). An assessment of the overall distributions of the ratings can be found in the second section of the appendix.

Table 1: Distribution of experiment participants by sex and German language competence.

	judg. type		
	A	**N**	**P**
German first lg.	40	38	38
of which female	31	29	32
German second lg.	4	3	5
of which female	3	3	4
n/a	1	0	1

The statistical analysis has been carried out with the statistical software R (R Core Team 2016). Nowadays, such data are usually analysed with linear mixed effects models (LMM), taking into account the interaction of fixed factors (like judgement type, sentence type, construction, grammaticality etc.) and random

25 Sociolinguistic findings on the usage of standard language in the English speaking and other societies have repeatedly uncovered the relevance of both sex/gender and social class in accounting for variation in the command and use of standard language (see a. o. Labov 1990; Chambers 2009: ch. 3; Labov 2006: ch. 8, 9). In a follow-up study to the present study, these aspects are explored in some detail (Vogel 2018).

factors (subject, item, lexical variant) with the goal of estimating the amount of variance that is explained by each factor. LMMs have become something like the state of the art in the analysis of experimental data. They have two major advantages over other methods: it is possible to deal with multiple random factors within one model and LMMs are quite robust against unbalanced data sets.

But the use of LMMs also draws attention away from effect sizes. Therefore, their use in our study is of limited relevance. The details of the calculations of LMMs that have been carried out are postponed to the appendix, fourth section.

The results of these calculations can be summarised easily, however: I first computed models evaluating the fixed factor *judgement type* for the five sentence types (unmarked, anglicisms, marked, gramatical taboos, ungrammatical) included in the experiment conditions and the filler sentences. For each of these sentence types, the fixed factor judgement type significantly improves the explanatory power of the LMM; with the exception of ungrammatical sentences, which differ only marginally between judgement types.

A second series of LMMs was calculated for each investigated grammatical taboo phenomenon. Here, the results are even more trivial: the best models are those that include the interaction of the fixed factors *grammaticality* and *taboo compliance*. That this is an absolutely obvious, but not very informative result, can be inspected from the results presented below.

To estimate confidence intervals of sentence types and the sizes of contrasts between them, I used parametric and non-parametric methods, reflecting the heteroscedasticity of the data. Confidence intervals and significance tests are based on t-tests with Welch approximation. These are used to estimate absolute values of empirical grammaticality and to test contrasts between them for significance.

The measure of effect sizes of contrasts that is used here is *Cliff's delta* (Cliff 1996), as already discussed in Section 2.2. Calculation of Cliff's delta has been undertaken with the *R* package *orddom* (Rogmann 2013).

As discussed above, some of the contrasts we are interested in (those comparing different levels of grammaticality) should produce at least medium effect size. The minimum sample size that is necessary to get a significant result with medium effect size for Cohen's *d* is 64 observations per condition.[26] This has been ensured in the experiment design. Comparisons of judgement types might produce smaller effect sizes, though. Likewise, comparisons of aggregated values per subject are slightly underpowered for this kind of comparison (48 subjects per judgement type).

26 Statistical power was calculated with the R package *pwr* (Champely 2017). Calculations are based on standard assumptions for type I errors (.05) and type II errors (.2).

Once a scale as in (2) is established for experimental morphosyntax, differences in mean acceptability can also directly be taken as effect size measures: for instance, the reasoning in Section 2 and the data to be discussed below suggest that a difference in mean acceptability of more than 0.25 is presumably a categorical difference, and a difference in mean acceptability of less than 0.10 is very likely negligible for linguistic theory.

4.3 Statistical analysis of judgement types, sentence types and taboos

Table 2 displays the mean ratings for the different sentence types included in this study. The set of "unmarked sentences" in Table 2 includes the test sentences in the unmarked conditions (+gramm/+taboo) of our four taboo phenomena, here collapsed into one category. The category "anglicsim" covers the filler sentences which are morphosyntactically unmarked, but contain English loan words. The filler sentences with "marked" status are summarised under that label. The same holds, accordingly, for the ungrammatical filler sentences. The four taboo phenomena are collapsed into one category for this examination.

Table 2: Mean ratings for the investigated sentence types under the three judgement types.

M	A	N	P	category (H II)
unmarked test sentences (collapsed)	0.811	0.841	0.928	✓
anglicism filler sentences	0.534	0.603	0.797	?
marked filler sentences	0.278	0.307	0.422	??
gr. taboo phenomena (collapsed)	0.187	0.245	0.361	??
ungrammatical filler sentences	0.085	0.097	0.093	*

The final column of Table 2 reports the range of acceptability to which the means of the sentence types belong under judgement type P. The outcome is as predicted in hypothesis H II (2) which can be seen as confirmed, to the extent that such a use of rating scales is sufficiently reliable.

The four taboo phenomena, taken together, also fall into the category of marked sentences under judgement type P, as expected. Ratings under judgement types N and A are worse than for P across the board. Judgement type A ratings are only slightly worse than those for judgement type N. This might suggest that aesthetic and normative judgements are not independent from each other. The comparatively largest, but still small, differences between judgement types A and N are found

Table 3: Standard deviations for the investigated sentence types under the three judgement types.

SD	A	N	P
unmarked test sentences (collapsed)	0.224	0.215	0.163
anglicism filler sentences	0.266	0.259	0.251
marked filler sentences	0.306	0.315	0.364
gr. taboo phenomena (collapsed)	0.218	0.249	0.337
ungrammatical filler sentences	0.161	0.165	0.183

with the two stigmatised phenomena: anglicisms and grammatical taboos which, it seems, have an additional aesthetic disadvantage.

Table 3 displays the standard deviations for the ratings of our five sentence types in all three judgement types. We envisage a consistent pattern: the closer the mean is to the centre of the scale (i. e., 0.5, see Table 2), the larger is the standard deviation. Grey shading in Table 3 highlights the highest values in each row. This is also the cell where the mean is closest to 0.5 for each row in Table 2 (except for ungrammatical sentences, but here the ratings differ only very marginally anyway). This results from a well-known correlation of means and their standard deviations: in an unskewed data set, the possible range for deviation from the mean becomes smaller, the closer the mean is to the lower or higher limit of the scale.

To estimate the effect size of judgement type for the different sentence types, I respected the between-subject heterogeneity in the following way: for each sentence type, the subjects' means have been aggregated. Pairwise comparisons of subjects' means between the three judgement types for each sentence type have then been carried out, calculating confidence intervals, Cliff's delta for effect size and t-tests with Welch approximation. Table 4 summarises the results for the differences between the two extremes, the judgement types P and A for each sentence type.

Grey shading of the Cliff's delta estimates signals large (dark grey) and medium (medium grey) effect sizes. All contrasts are significant in a t-test except for the ungrammatical sentences which are rated very low overall. Grammatical taboos yield a higher contrast than marked sentences, and likewise anglicisms yield a higher contrast than unmarked sentences, although their rating under judgement type P is lower. This is due to the disadvantage of the stigmatised constructions under the aesthetic judgement type A.

Hypothesis H I postulates at least medium effect sizes for pairwise comparisons of our four categories of acceptability. To test this hypothesis, I calculated pairwise comparisons under judgement type P (with anglicisms counting as slightly marked

Table 4: Pairwise comparisons of subjects' means for the judgement types P and A for each sentence type (Bonferroni-corrected for three pairwise comparisons). Mean difference, Cliff's delta (both with 95% confidence intervals) and Welch's t-test.

P – A	estimate	Cliff's delta	t-ratio	df	p-value
unmarked	0.121 ±0.059	0.533 [0.252; 0.732]	5.291	87	8.98e-07 ***
anglicisms	0.262 ±0.097	0.661 [0.404; 0.821]	7.659	87	2.43e-11 ***
marked	0.144 ±0.089	0.438 [0.150; 0.658]	4.077	87	0.0001 ***
gr. taboos	0.177 ±0.094	0.487 [0.204; 0.695]	4.703	87	9.56e-06 ***
ungrammatical	0.007 ±0.043	0.121 [−0.177; 0.400]	0.984	87	0.3276

Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

Table 5: Four sentence types under judgement type P. Results for pairwise comparisons of subjects' means: point estimates and 95% confidence intervals of mean rating differences, paired Cliff's delta with 95% confidence intervals, Welch's t-tests (Bonferroni-corrected for six comparisons).

contrast	estimate	Cliff's delta	t-ratio	df	p-value
unm. – angl.	0.132 ± 0.064	0.682 [0.276; 0.881]	6.111	43	2.51e-07 ***
angl. – mark.	0.374 ± 0.076	0.909 [0.560; 0.984]	14.310	43	0 ***
mark. – ungr.	0.331 ± 0.075	1 [0.704; 1]	inf.	43	0 ***

Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

and leaving out grammatical taboos for the moment, see Table 5). Only the three crucial pairwise comparisons are reported in Table 5.[27] The calculation of Cliff's delta in Table 5 only takes into account the value pairs produced by each subject.[28] The maximum value of 1 in the comparison of marked and ungrammatical sentences is due to the fact that all subjects gave the higher rating to the same sentence type. Overall, the contrasts have large effect sizes.

In other words, the categorical distinctions of the four levels of grammaticality assumed by the experts are very consistently confirmed by the participants under judgement type P. So hypothesis H I, just like hypothesis H II, can be seen as largely confirmed by the results of this study. We will now take a closer look at grammatical taboos.

27 The three missing comparisons have an effect size of d = 1.

28 This within-subject analysis is only possible when analysing ratings under the same judgement type. It differs from the between-subjects analysis that is necessary when comparing different judgement types, where all possible pairs of subjects' means from the two groups have to be used.

4.4 Grammatical taboos and grammar-internal markedness

Can grammatical taboos be distinguished empirically from grammar-internal markedness? We have already seen that grammatical taboos seem to have a slight aesthetic disadvantage under judgement type A. But all the statistics tells us up to here, is that our grammatical taboos, just like the marked filler sentences, fall into the range of markedness that we defined earlier, especially under judgement type P, which is the crucial judgement type for the linguistic analysis.

I will now examine the distributions of the judgements for these two groups more closely in a number of post hoc analyses. We will see that the two samples differ to some extent in their patterns of variation. Interestingly, this only shows up when we consider the variation among subjects under judgement type P. Compare the very similar distributions of all observed ratings for marked sentences and grammatical taboos in Figure 2 with the distributions of subjects' mean ratings for these sentence types in Figure 3.[29]

While the distribution for marked sentences in Figure 3 has a peak at level "–1", the distribution for grammatical taboos is flatter and wider with a plateau enclosing the levels "–2" to "0".[30] One measure for quantifying this visual impression is the

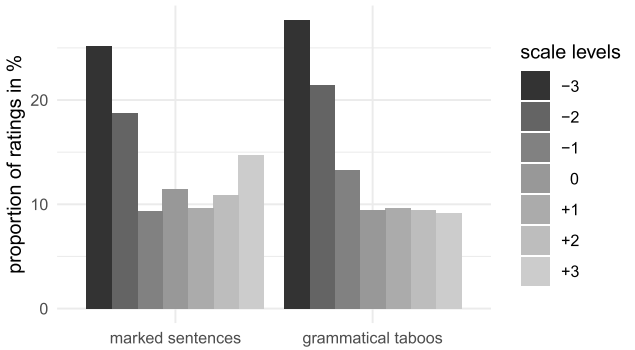

Figure 2: Proportional distribution of ratings over the seven scale levels for marked sentences (N = 394) and grammatical taboos (N = 383), judgement type P.

29 Each subjects' mean for grammatical taboos is computed as mean over her means for each of the four taboo phenomena. This ensures that each taboo phenomenon contributes with equal weight, even when some values might be missing. For the sake of completeness: the variance patterns of subjects' ratings of grammatical taboos under the judgement types N and A differ from P and are similar to that of marked sentences in Figure 3.

30 We furthermore do not observe a bimodal distribution as in the experiment by Bader and Schmid (2006) (see again Figure 1).

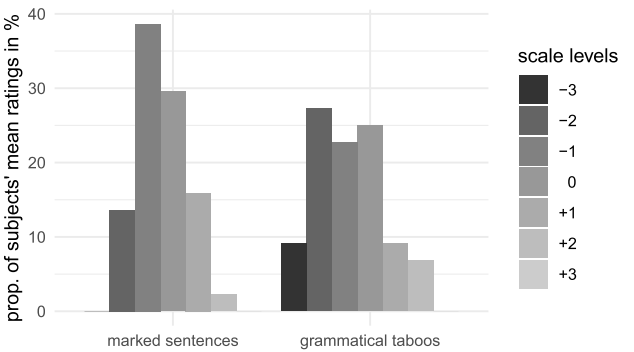

Figure 3: Proportional distribution of subjects' mean ratings (N = 44) over the seven scale levels for marked sentences and grammatical taboos, judgement type P.

Table 6: Excess kurtosis for the four distributions in Figures 2 and 3.

	by obs. (Fig. 2)	by subj. (Fig. 3)	diff.
marked filler sentences	−1.351	−0.344	1.007
grammatical taboos (collapsed)	-0.994	-0.736	0.258

kurtosis which has been introduced by Karl Pearson and measures the amount of outliers in a distribution (its "tailedness") in comparison to a normal distribution. Table 6 displays the (excess kurtosis) values for the four distributions illustrated in Figures 2 and 3.[31]

When we aggregate plain observations by subject, we expect the amount of outliers to be reduced, because the influence of one random factor (the items for the marked fillers, and the different taboo phenomena for grammatical taboos) has been neutralised. This is indeed the case for our marked fillers where kurtosis for the by subjects sample is reduced by 1.007 and now much closer to 0, the kurtosis of a normal distribution. But in the case of grammatical taboos, reduction of negative kurtosis is quite small. A post hoc Levene-test for homogeneity of variances for the two by-subjects samples confirmed that the null hypothesis of variance homogeneity can be rejected (df = 1, 86; F= 6.6032, p = 0.0119 *).

31 The negative values indicate a kurtosis that is smaller than in a normal distribution which has an excess kurtosis of 0. The simple kurtosis of a normal distribution is 3. Excess kurtosis is the simple kurtosis subtracted by 3. I am using excess kurtosis for ease of exposition only. The calculation of kurtosis has been carried out with the R package *moments* (Komsta and Novomestky 2015).

For the calculation of kurtosis, outliers are those data points that are outside of the range of M ± SD (Westfall 2014). A larger SD therefore tends to reduce the number of outliers in this sense. This seems to be the underlying cause for the observation in Table 6.

As we discussed in connection with Table 3, the SD is influenced by the mean. An often used measure to control for this is the *coefficient of variation* (CV) which is the standard deviation divided by the mean. Table 7 shows means, standard deviations and CVs for the two by subjects samples.

Table 7: Means, standard deviations and coefficients of variation for subjects' means of marked filler sentences and grammatical taboos (collapsed), judgement type P.

	M	SD	CV
marked filler sentences	0.423	0.166	0.392
grammatical taboos (collapsed)	0.358	0.219	0.610

The SD of grammatical taboos is higher, although their mean is a bit lower than that of the marked fillers. This results in a CV for grammatical taboos that is about 56 % higher than that of the marked filler sentences (0.610/0.392 = 1.556). A post hoc test for equality of CVs, following the method proposed by Feltz and Miller (1996), confirmed that the null hypothesis of equal CVs can be rejected (D-AD = 5.409; p = 0.020 *).[32]

As discussed in Section 2, tests for significance alone are usually not very informative. This also holds here. The crucial question is whether the *quantitative* difference of a 56 % higher CV for grammatical taboos justifies the assumption of a *qualitative* contrast between the two types of phenomena. While this might not be unreasonable, the idea of *differences in variances* as another type of effect to look at in addition to the usually analysed *differences in means* thus far has been explored very rarely, if at all.[33]

A further cause of variation in the experiment is judgement type. Table 8 shows that CVs for the marked filler sentences decrease from type A via N to P. The contrast is similar for grammatical taboos with type A having the highest CV, but there is no reduction of the CV for judgement type P, it is even slightly larger than for type N.

32 The test has been carried out with the R package *cvequality* (Marwick and Krishnamoorthy 2018).

33 An important methodical problem that needs to be solved in such an approach is that standardised effect sizes for *contrasts between variances as an effect* have not yet been established.

Table 8: CVs for marked sentences and grammatical taboo phenomena, aggregated over subjects' means, for the three judgement types.

CV	A	N	P
marked filler sentences	0.632	0.492	0.392
gr. taboo phenomena (collapsed)	0.734	0.589	0.610
CV_{taboo}/CV_{marked}	1.161	1.197	1.556

This suggests that our non-prescriptive judgement type P only makes the rating of ordinary marked sentences easier, but not the rating of grammatical taboos.

Judgement type P seems to focus on the *paradox* of grammatical taboos, without being able to neutralise stigmatisation: participants are requested to decide between the two equally salient contradictory options that make up the paradox.[34]

The participants were drawn from a homogeneous population: students of nearly the same age mostly from the same local region who just finished school and started their BA program – mostly with the aim of becoming a school teacher for German studies. They made similar experiences with the German standard language during their 12 or 13 school years. It might therefore be surprising to find such disagreement. In Vogel (2018), I present follow-up studies in which I am trying to get a grip on sociolinguistic factors that may correlate with the inter-individual differences within this (not so?) homogeneous group.

4.5 Comparison of grammatical taboos

This section inspects the four taboo phenomena in more detail. A summary statistics for judgement type P is given in Table 9. We see some differences in the CV values of the four taboo phenomena which seem to be correlated with salience. Auxiliary *tun* and *weil* V2-clauses have a higher CV than the two non-salient phenomena, i. e. they seem to be more controversial among subjects due to their salience.

Figure 4 displays the mean ratings with confidence intervals for our four grammatical taboos under judgement type P. While the three other phenomena are solidly settled in the area of markedness, auxiliary *tun* is at its lower margin with the confidence interval reaching below 0.20, leaving a little bit of doubt about its grammaticality status. In the following analyses, auxiliary *tun* will therefore be

34 Under this interpretation, our findings can be seen as confirmation of the observation by Bader and Schmid (2006), discussed above (see Figure 1).

Table 9: Result statistics over subjects' means for each of the four grammatical taboos, and the marked filler sentences: means (M), standard deviations (SD) and coefficients of variation (CV); judgement type P.

gramm. taboo	M	SD	CV
aux. *tun*	0.254	0.269	1.060
weil V2-clause	0.375	0.327	0.873
d-pronoun	0.396	0.296	0.748
double perfect	0.408	0.266	0.651
marked filler sentences	0.423	0.166	0.392

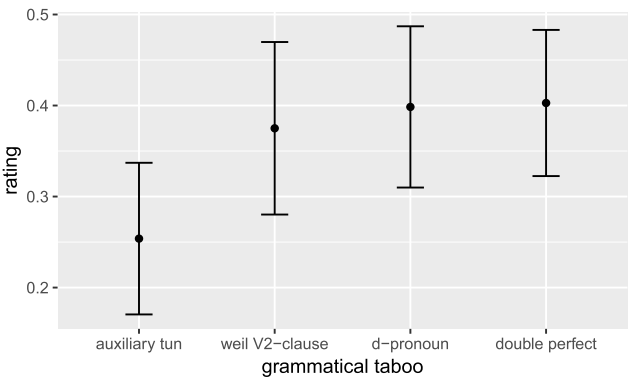

Figure 4: Means and 95% confidence intervals (Bonferroni-corrected) for the four grammatical taboo phenomena; judgement type P.

of particular interest. Table 10 summarises the results of pairwise comparisons of auxiliary *tun* with the other three phenomena under judgement type P. In Figure 4, auxiliary *tun* is separated from the other three phenomena in absolute acceptability. Our non-parametric effect size measure relativises this for the second salient phenomenon, *weil* V2-clause. Comparison of auxiliary *tun* and *weil* V2-clause produces a non-significant contrast with Cliff's delta of only small effect size. This might signal a non-categorical contrast. This is in line with our assumptions. For the contrast between auxiliary *tun* and the two non-salient taboos, we find medium and large effect sizes, signalling the expected categorical contrast.[35]

35 The *weil* V2-clause condition contrasts with neither of the three other phenomena to a sufficient degree, which also suggests an intermediate status.

Table 10: Grammatical taboo phenomena under judgement type P. Pairwise comparisons of subjects' means for auxiliary *tun*: point estimates and 95% confidence intervals of mean rating differences, paired Cliff's delta with confidence intervals, Welch's t-tests (conf. intervals and sign.-level indicators Bonferroni-corrected for six comparisons).

aux. *tun* vs ...	estimate	Cliff's delta	t-ratio	df	p-value
weil V2-clause	0.121 ±0.153	0.205 [−0.164; 0.523]	1.545	43	0.130
d-pronoun	0.142 ±0.117	0.386 [0.006; 0.669]	3.046	43	0.0039 *
double perfect	0.154 ±0.110	0.477 [0.089; 0.739]	3.857	43	0.00038 **

Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

Table 11: Means of subjects' 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.

M	A	N	P	P − A
aux. *tun*	0.137	0.165	0.254	0.117
weil V2-clause	0.144	0.260	0.375	0.231
d-pronoun	0.219	0.293	0.396	0.177
double perfect	0.228	0.269	0.408	0.180

Our hypotheses overall bear on *two scale effects* (judgement type and salience of the taboo phenomena) and the *contrast* between taboo violations and ungrammaticality (it should have a size of Cliff's delta > 0.333). We start with an overview of the means of each grammatical taboo phenomenon under the three judgement types in Table 11. The effect of the two scales comes out quite clearly: for each cell (ignoring the rightmost column) in Table 11 the cells downwards and rightwards have higher values – with one irrelevant exception (grey shaded).

The rightmost column in Table 11 reports the differences for judgement types P and A. It provides some hints as to the effect size of judgement type. Interestingly, the two non-salient phenomena are here in the middle, whereas auxiliary *tun* has the lowest value – due to low rating under judgement type P – and the other salient phenomenon, *weil* V2-clause, is highest. It has a rating as high as the non-salient phenomena for types P and N, but a low rating at the level of auxiliary *tun* for type A, which leads to the highest difference between types P and A. This suggests a gradient difference among the salient taboos such that *weil* V2-clauses are degraded specifically under the aesthetic perspective, whereas auxiliary *tun* is stigmatised strongly across the board.

Table 12: Pairwise comparisons of *subjects' means* for the judgement types P and A for each grammatical taboo (Bonferroni-corrected for three comparisons). Mean difference, Cliff's delta and Welch's t-test (sign.-indicators Bonferroni-corrected for three comparisons).

P – A	estimate	Cliff's delta	t-ratio	df	p-value
auxiliary *tun*	0.117 ±0.119	0.271 [–0.019; 0.519]	2.369	87	0.02
weil V2-clause	0.231 ±0.108	0.419 [0.186; 0.607]	3.866	87	0.0002 ***
d-pronoun	0.177 ±0.129	0.351 [0.057; 0.589]	3.103	87	0.0026 **
double perfect	0.180 ±0.120	0.391 [0.097; 0.623]	3.512	87	0.0007 **
marked	0.144 ±0.089	0.438 [0.150; 0.658]	4.077	87	0.0001 ***
ungrammatical	0.007 ±0.043	0.121 [–0.177; 0.400]	0.984	87	0.3276

Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

The contrasts between subjects' aggregated mean ratings under judgement types P and A for the four taboo phenomena are displayed in Table 12. The values for the marked and ungrammatical fillers from Table 4 are repeated for comparison.

Auxiliary *tun* produces only a small effect size. The contrast does not reach statistical significance under Bonferroni-correction (this would require $p < .01667$). The main reason is that auxiliary *tun* has a quite low rating already under judgement type P. Still, effect size of auxiliary *tun* is higher than for ungrammatical sentences, though lower than for ordinary markedness. This corroborates the impression that it is a very strong case of markedness. The other three taboo phenomena show effect sizes similar to ordinary markedness. This is in part due to the aesthetic stigma associated with even weaker taboos, resulting in lower rating under judgement type A than for ordinary marked sentences. It can therefore be interpreted as one of the empirical effects of the general standard language taboo.

We will now turn to the subexperiments for the grammatical taboos as described in Section 4.1. Recall that for each of the four investigated grammatical taboos, four conditions were constructed in a 2x2 design.[36] The two factors relevant for the analysis are *taboo compliance* and *grammaticality*. Random factors are *subject* and *lexical variant*. Each of the four taboo phenomena was investigated under three different judgement types, so that overall we have 12 subexperiments.

36 The initial 3x2 design of the double perfect was reduced to a 2x2 design by conflating double perfect with double past perfect, justified by marginal differences in means:

	A	N	P
double perfect	0.23	0.27	0.40
double past perfect	0.23	0.24	0.41

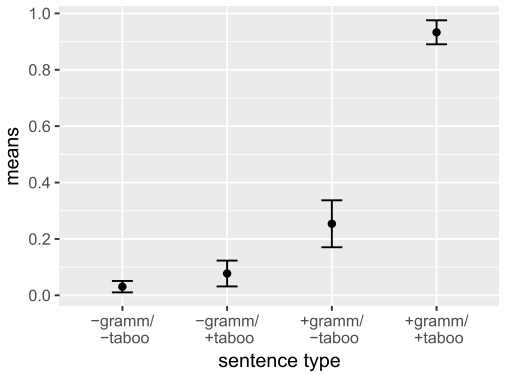

Figure 5: Means and 95% confidence intervals (Bonferroni-corrected) for test sentences with(out) auxiliary *tun* under judgement type P.

The discussion is limited to the most relevant aspects with special emphasis on judgement type P.

The relevance of the interaction of the two fixed factors is witnessed by the great distance in acceptability between the +gramm/+taboo condition and the three others. This is exemplified for auxiliary *tun* in Figure 5.

The tendencies that we see in Figure 5 are by and large the same with the other judgement types and taboo phenomena. Pairwise comparisons of contrasts between the four conditions in each of the four subexperiments for judgement type P led to significant results in all cases except for the two ungrammatical conditions. Under judgement types A and N, contrasts between taboo violations and ungrammatical sentences become smaller. These judgement types focus on the prescriptive side of the paradox of grammatical taboos and thus favour a resolution of the paradox in the direction of non-acceptability.

We see in Figure 5 that the 95% confidence interval for auxiliary *tun*, the +gramm/−taboo condition, ranges below 0.20, our predefined minimum for marked sentences, but only slightly so. Remember that this is only a rule of thumb. We could even deliberately lift this level up to 0.30, but there would be another price to pay: admittance of gradient ungrammaticality. This becomes clear from the data after inspecting the contrast that interests us most: the contrast between a taboo violation (+gramm/−taboo) and a grammaticality violation (−gramm/+taboo). Table 13 summarises the results from the pairwise comparisons in all 12 subexperiments for the contrast between these two conditions. Indicated are the effect sizes of the contrasts (Cliff's delta) and the significance levels of the pairwise comparisons (Welch's t-test, Bonferroni-corrected for six comparisons).

Table 13: Pairwise comparisons of subjects' means for the +gramm/−taboo and −gramm/+taboo conditions for all taboo phenomena and judgement types: paired Cliff's delta and (Bonferroni-corrected) indicators of significance level for Welch's t-tests.

	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 ***

Cliff's delta: small = d > 0.147; medium = d > 0.333; large = d > 0.474

The grey shaded cells in Table 13 are those, where Cliff's delta has medium or large effect size which we can safely assume to signal a categorical difference between the taboo phenomena and their ungrammatical counterpart without taboo violation.

Auxiliary *tun* contrasts with the ungrammatical condition to a smaller extent than the other phenomena, but still with medium effect size even under judgement type N. The salience of the taboo phenomena comes out quite clearly under judgement type A, where Cliff's delta is below the threshold even for small effect size, whereas the non-salient phenomena produce a large effect size across the board.

Only three cells do not show a large effect size, the two salient phenomena under judgement type A and auxiliary *tun* under judgement type N. These are also those cases where the mean is below 0.2 (see Table 11 again). This parallel conforms with the expectations formulated in our hypotheses H I and H II.

Judgement type P opens the space for the taboos at least in that here even auxiliary *tun* contrasts with the ungrammatical condition with highly significant large effect size. It therefore can safely be classified as marked, alongside with the other phenomena. Another interesting aspect is the fact that even under the two prescriptive judgement types participants appear to be quite tolerant: the taboo phenomena are not conflated with ungrammatical sentences under judgement type N, as we would expect from the treatment of these phenomena in reference grammars (Duden 2009; Duden 2016). To yield such a conflation, it takes a stronger perspective like the aesthetic one, but even here it only works for the two salient taboo phenomena.

While it seems to be impossible to totally neutralise the effects of stigmatisation even under the very tolerant judgement type P, as witnessed by the marked ratings for grammatical taboos, it is also true that the prescriptive judgement types A and N cannot enforce the kind of black-and-white attitude towards informal language that is typical of the prescriptive literature. Non-salient taboos seem to remain below the prescriptive radar in such an experimental setting.

So we can conclude that the participants of the experiment accept the more restrictive attitude of normative grammatical thinking in principle, but apply it moderately in their ratings.

5 Summary and outlook

Non-experimental morphosyntax, as it has been practised for decades, is relying on the *construction* of the facts of the target language on the basis of expert opinion and expert consent – be it the expert knowledge of linguists themselves about their mother tongue, or field work with few, often just one informant serving as expert(s) for a particular language – supported, wherever possible, by corpus evidence.

The experimental turn that we currently envisage is motivated by scepticism about the reliability of this approach. Though it could be shown that scepticism is often unjustified, there are two obvious limitations which we hope to overcome using experimental methods, both of which have to do with the limitations of the individual expert. Some phenomena may only be detectable via experimental methods, either because they concern effects which are too small to identify without careful experimental investigation, or because they have sociocultural causes which usually are invisible to the individual researcher or speaker who rarely is able to see beyond her own social class/group from the armchair.

It seems reasonable to me, not only for the present study, to follow the de Saussurean/sociolinguistic view of grammar as a social entity. It develops over time within a community that constantly reorganises the language it shares by means of a kind of contract its members implicitly agree on. As much as there is inequality within communities in socioeconomic or cultural terms, there is also inequality of speakers: some (groups of) speakers are more privileged than others in the options they have to shape the system of the standard language and the community's ideological dispositions – and thereby indirectly also influence the direction of language change. Grammatical taboos, i. e. the stigmatisation of particular constructions which are often used by less privileged groups or for less prestigious purposes, belong to these ideological dispositions and can frequently be observed.

This is especially important when we focus on *standard languages*. And the languages practised most often in larger societies are more or less standard varieties. For the major part of the German speaking community, the standard language has even become the mother tongue, but in an informal register with interesting divergences from the written standard which are subject to the GSLT.

But grammatical theory is interested in the grammars of *natural* languages. How "natural", after all, are languages actually, especially standard languages?

What do we mean by "natural"? And how does the linguist, just from the armchair, come to recognise and sort out the artificial properties of a standard language?

With respect to auxiliary *tun*, perhaps most German linguists can agree that the informal registers of German, even standard German, do not exclude this construction. But this does not mean that the theory of the phenomenon is free of ideas typical of the standard language ideology.

Consider the constraint "Full Interpretation" (FI) which originated from Chomsky's (1986; 1995) work and formulates the requirement that morphosyntactic units must be interpreted, i. e. enter the "interfaces" to logical and phonetic form non-empty.[37] Language, according to this principle, is redundancy-free. This view has a certain parallel in the prescriptive discourse, where the use of allegedly superfluous expressions is frequently dismissed as bad language. Grimshaw (1997) makes crucial use of her own version of FI in her analysis of English *do*-support. It is also used by Bader and Schmid (2006) in their analysis of German auxiliary *tun*:

(18) **Full Interpretation (FULLINT)**
Lexical conceptual structure is parsed. (Grimshaw 1997)

This constraint, different in formulation but not in spirit from Chomsky's initial proposal, effectively penalises the use of function words, and in particular the grammaticalisation of function words from lexical words. The idea is that grammatical uses of verbs like English *have, be, do* etc. require ignoring their lexical meaning, e. g. the auxiliary use of *have* in the perfect construction does not entail the possessive semantics this verb has in its lexical use.

As described in Section 3, it is part of the *complaint tradition* of the German standard language ideology that use of function words and analytic inflection are stigmatised as signs of *language decline*. I don't see much difference between this ideological opinion on grammar and the constraint in (18). With FULLINT as a typological universal (due to its status as constraint in an OT analysis), this stigmatisation is given the same kind of pseudo-naturalistic legitimisation that can be found in the prescriptive literature.

But on the other hand, ironically: if we understand FULLINT as an ideological constraint, we reconstruct the finding that auxiliary *tun* is ruled out on ideological grounds – as it in fact happened (cf. Langer 2001; Davies and Langer 2006).[38]

37 As is often the case with Chomsky's work, FI can be understood, a. o., as one particular constraint in the grammar, an architectural design feature of grammar, or as part of the definition of "morphosyntactic unit".

38 An analysis of the restrictions on English *do*-support that avoids this trap is sketched in Vogel (2013). There, both variants with and without auxiliary *do* are assumed as well-formed, but

This may illustrate the risks of armchair linguistics, and the fallacy of prescriptive bias which linguists like everyone can easily fall victim to. Experimental methods might help to avoid this. Our study also showed that grammatical taboos (GT) in general fall in the range of markedness with respect to empirical grammaticality. They differ from ordinary grammar-internal markedness (OM) in several ways:

i) We have indications that the variance between subjects under judgement type P is substantially larger for GT than for OM. This outcome is expected under the assumption that stigmatisation (having an ideological basis) is followed less consistently by speakers than markedness that is due to grammar-internal factors.
ii) Grammatical taboos have a particular disadvantage under the aesthetic judgement type A which is larger for the salient taboos. Judgement type A directs subjects towards the prescriptive resolution of the paradox of GTs. Put differently: grammatical taboos, as discussed in the prescriptive literature, often come with an aesthetic devaluation which can be uncovered using an aesthetically oriented judgement type.
iii) The method of using different judgement types also uncovered the difference between the salient taboos auxiliary *tun* and *weil* V2-clause, whereby with respect to our numeric measures of empirical grammaticality only under the aesthetic perspective *weil* V2-clauses grouped together with auxiliary *tun*. Both were no longer distinguishable from ungrammaticality under judgement type A.

Even the strongest taboo examined here, auxiliary *tun*, is grammatical, according to the measures applied in this study which posit the minimum acceptability of grammatical sentences at 0.20 under judgement type P.

We tried to follow closely the advice from the "new statistics" (Kline 2013), focusing on point estimates, their confidence intervals and effect sizes of contrasts, downgrading the role of p-values, testing substantial hypotheses rather than nil hypotheses, and using *estimation* language. In order to be able to do this, we had to introduce the concept of empirical grammaticality and formulate ranges of

---

the variant with *do*-support, being the larger construction, triggers an implicature that leads to pragmatic enrichment, in this case verum focus (as in *I DO like it*.), whereas it is pragmatically blocked by the simple present tense version under normal conditions.

For German, the situation is even clearer as the grammar of verbal inflection in German shows a strong tendency towards analytic inflection in most cases. So, auxiliary *tun* looks more like the default – which underpins the assessment that its exclusion cannot have a grammar-internal foundation.

acceptability that correspond to the levels of gradient grammaticality assumed in the theoretical literature.

In this respect the study is only a proof of concept. It seems to work, by and large, for elicitation of Standard German with rating scales in a written questionnaire with university students. Follow-up research and meta-analytical studies on different languages, designs and participant groups are necessary to further substantiate the feasibility of such an approach.

One particular question with respect to comparing different methods has already come up when we compared the distribution of subjects' means in the present study (Figure 3), which showed a flat and wide distribution, with the observation of a bimodal distribution for auxiliary *tun* by Bader and Schmid (2006) reported in Figure 1. The hypothesis to be checked for is that this difference in the character of the distributions is due to the experimental task: for grammatical taboos, gradient rating scales lead to a flat and wide distribution, whereas a binary yes/no decision task leads to a bimodal distribution.

Acknowledgment: The research and analyses presented in this paper have grown over several years. This work has been presented in different versions at different stages of the development of this paper at the universities of Bielefeld and Leipzig, the Minsk State Linguistic University, Belarus, and the University of Wuppertal. I want to thank the audiences of these presentations for very helpful comments and suggestions. My special thanks goes to my student assistants, first of all Ann-Christin Broschinski, without whom empirical studies like the one presented here could not be undertaken. I also want to thank two anonymous reviewers and the editors of ZS, especially Gerhard Jäger, for very helpful comments and suggestions.

References

Antomo, Mailin & Markus Steinbach. 2010. Desintegration und Interpretation: *Weil*-V2-Sätze an der Schnittstelle zwischen Syntax, Semantik und Pragmatik. *Zeitschrift für Sprachwissenschaft* 29. 1–37.

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46. 273–330.

Bader, Markus & Tanja Schmid. 2006. An OT-analysis of *do*-support in Modern German. Online document. Manuscript. http://roa.rutgers.edu/files/837-0606/837-BADER-0-0.PDF (06.02.2019).

Barbour, Stephen & Patrick Stevenson. 1998. *Variation im Deutschen*. Berlin & New York: de Gruyter.

Bosch, Peter, Graham Katz & Carla Umbach. 2007. The non-subject bias of German demonstrative pronouns. In Monika Schwarz-Friesel, Manfred Consten & Mareile Knees (eds.), *Anaphors*

in text. Cognitive, formal, and applied approaches to anaphoric reference, 145–164. Amsterdam: Benjamins.
Bosch, Peter & Carla Umbach. 2007. Reference determination for demonstrative pronouns. In Dagmar Bittner & Natalia Gargarina (eds.), *Intersentential pronominal reference in child and adult language* (ZAS Papers in Linguistics 48), 39–51. Berlin: Zentrum für Allgemeine Sprachwissenschaft.
Chambers, Jack K. 2009. *Sociolinguistic theory*. Revised edn. Chichester: Wiley-Blackwell.
Champely, Stephane. 2017. *pwr: Basic functions for power analysis*. R package version 1.2-1. https://CRAN.R-project.org/package=pwr (06.02.2019).
Chomsky, Noam. 1986. *Knowledge of language. Its nature, origin and use*. New York: Praeger.
Chomsky, Noam. 1995. *The minimalist program*. Cambridge, MA: MIT Press.
Cliff, Norman. 1996. *Ordinal methods for behavioral data analysis*. London: Routledge.
Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ: Erlbaum.
Cohen, Jacob. 1992. A power primer. *Psychological Bulletin* 112(1). 155–159.
Cohen, Jacob. 1994. The earth is round (p < .05). *American Psychologist* 49(12). 997–1003.
Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.
Davies, Winifred & Nils Langer. 2006. *The making of bad language*. Frankfurt/Main a.o.: Peter Lang.
Devitt, Michael. 2006. *Ignorance of language*. Oxford: Clarendon Press.
Duden 2009. *Duden. Die Grammatik (Duden 4)*. 8th edn. Mannheim [a. o.]: Dudenverlag.
Duden 2016. *Duden. Die Grammatik (Duden 4)*. 9th edn. Berlin: Dudenverlag.
Durrell, Martin. 1999. Standardsprache in Deutschland und England. *Zeitschrift für Germanistische Linguistik* 27(3). 285–308.
Eisenberg, Peter. 1986. *Grundriss der deutschen Grammatik*. Stuttgart: Metzler.
Elspaß, Stephan. 2005. Standardisierung des Deutschen. Ansichten aus der neueren Sprachgeschichte 'von unten'. In Ludwig M. Eichinger & Werner Kallmeyer (eds.), *Standardvariation. Wie viel Variation verträgt die deutsche Sprache?* (Jahrbuch des IDS 2004), 64–99. Berlin & New York: de Gruyter.
Elspaß, Stephan. 2010. Klammerstrukturen in nähesprachlichen Texten des 19. und frühen 20. Jahrhunderts. In Arne Ziegler (ed.), *Historische Textgrammatik und historische Syntax des Deutschen*, 1011–1026. Berlin & New York: de Gruyter.
Featherston, Sam. 2005. The decathlon model: Design features for an empirical syntax. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187–208. Berlin & New York: de Gruyter Mouton.
Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33(3). 269–318.
Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28(1). 127–132.
Feltz, Carol J. & Gerald E. Miller. 1996. An asymptotic test for the equality of coefficients of variation from k populations. *Statistics in Medicine* 15(6). 646–658.
Gigerenzer, Gerd, Stefan Krauss & Oliver Vitouch. 2004. The null ritual. What you always wanted to know about significance testing but were afraid to ask. In David Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences*, 391–408. Thousand Oaks, CA: Sage Publications.
Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry* 28(3). 373–422.

Hubers, Ferdy, Tineke Snijders & Helen de Hoop. 2016. How the brain processes violations of the grammatical norm: An fMRI study. *Brain and Language* 163. 22–31.

Kline, Rex B. 2013. *Beyond significance testing: Statistics reform in the behavioral sciences*. Washington, DC: American Psychological Association.

Komsta, Lukasz & Frederick Novomestky. 2015. *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14. https://CRAN.R-project.org/package=moments (06.02.2019).

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

Labov, William. 2006. *Principles of linguistic change. Volume 2: Social factors*. Chichester: Wiley-Blackwell.

Labov, William. 2010. *Principles of linguistic change. Volume 3: Cognitive and cultural factors*. Chichester: Wiley-Blackwell.

Langer, Nils. 2001. *Linguistic purism in action. How auxiliary* tun *was stigmatised in Early New High German*. Berlin & New York: de Gruyter.

Marwick, Ben & Kalimuthu Krishnamoorthy. 2018. *cvequality: Tests for the equality of coefficients of variation from multiple groups*. R package version 0.1.3. https://github.com/benmarwick/cvequality (06.02.2019).

McBryde, John M. 1943. Some grammatical taboos. *Peabody Journal of Education* 20(5). 272–279.

Milroy, James & Lesley Milroy. 1985. *Authority in language*. London: Routledge & Kegan Paul.

von Polenz, Peter. 1994. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Bd. II: 17.-18. Jahrhundert*. Berlin & New York: de Gruyter.

von Polenz, Peter. 1999. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Bd. III: 19.-20. Jahrhundert*. Berlin & New York: de Gruyter.

von Polenz, Peter. 2000. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Bd. I: Einführung, Grundbegriffe, 14.-16. Jahrhundert. 2nd rev. & enl. edn*. Berlin & New York: de Gruyter.

Portele, Yvonne & Markus Bader. 2016. Accessibility and referential choice: Personal pronouns and d-pronouns in written german. *Discours* [Online] 18/2016. 1–39. DOI: 10.4000/discours.9188.

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/ (06.02.2019).

Rogmann, Jens J. 2013. *orddom: Ordinal dominance statistics*. R package version 3.1. https://CRAN.R-project.org/package=orddom (06.02.2019).

Romano, Janine, Jeffrey D. Kromrey, Jesse Coraggio & Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the nsse and other surveys? Paper presented at the annual meeting of the Florida Association of Institutional Research, February 1–3, 2006, Cocoa Beach, FL.

de Saussure, Ferdinand. 1983 [1916]. *Course in general linguistics*. Edited by Charles Bailly and Albert Sechehaye with the Collaboration of Albert Riedlinger. Translated and annotated by Roy Harris. Chicago & La Salle, IL: Open Court.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: The University of Chicago Press.

Selting, Margret. 1999. Kontinuität und Wandel der Verbstellung von ahd. *wanta* bis gwd. *weil*. *Zeitschrift für Germanistische Linguistik* 27(2). 167–204.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87(2). 274–288.

Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics* 2(1). 14. DOI: 10.5334/gjgl.236.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134. 219–248.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.

Vogel, Ralf. 2013. The Trivial Generator. In Hans Broekhuis & Ralf Vogel (eds.), *Linguistic derivations and filtering. Minimalism and optimality theory*, 238–266. Sheffield: Equinox Publishing Ltd.

Vogel, Ralf. 2018. Sociocultural determinants of grammatical taboos in german. In Liudmila Liashchova (ed.), *The explicit and the implicit in language and speech*, 116–153. Newcastle upon Tyne: Cambridge Scholars Publishing.

Volodina, Anna. 2009. Epistemische Lesarten: konventionell oder pragmatisch? Online document. Slides from a conference presentation. http://www.annavolodina.de/dokumente/Vortrag_SPSW_Stuttgart_volodina.pdf (06.02.2019).

Weiß, Helmut. 2004. Zum linguistischen Status von Standardsprachen. In Maria Kozianka, Rosemarie Lühr & Susanne Zeilfelder (eds.), *Indogermanistik – Germanistik – Linguistik. Akten der Arbeitstagung der Indogermanistischen Gesellschaft, Jena, 18.-20.9.2002*, 591–643. Hamburg: Dr. Kovač.

Weiß, Helmut. 2005. Von den vier Lebensaltern einer Standardsprache – Zur Rolle von Spracherwerb und Medialität. *Deutsche Sprache* 4/2005. 289–307.

Weskott, Thomas & Gisbert Fanselow. 2009. Scaling issues in the measurement of linguistic acceptability. In Sam Featherston & Susanne Winkler (eds.), *The fruits of empirical linguistics I: Process*, 229–246. Berlin & New York: de Gruyter Mouton.

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273.

Westfall, Peter H. 2014. Kurtosis as peakedness, 1905–2014. r.i.p. *The American statistician* 68(3). 191–195.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache, Bd. I*. Berlin & New York: de Gruyter.

Zybatow, Tatjana & Thomas Weskott. 2018. Das Doppelperfekt: Theorie und Empirie. *Zeitschrift für Sprachwissenschaft* 37(1). 83–124.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/zfs-2019-0002). It contains information about the instructions used in the experiment, the overall distribution of the ratings in the experiment, the individual test sentences used as marked fillers and the linear mixed models that have been calculated.