Acceptability judgments in moribund heritage languages: Mitigating the challenges

Yvonne van Baal*

Abstract. Acceptability judgments are a frequently used method across linguistic disciplines, including heritage language linguistics. However, it has been argued that the method is not suitable for this population. This paper reviews the concerns that have been raised about acceptability judgments and presents two case studies that use judgment data from moribund Germanic heritage languages. These illustrate the strengths and obstacles of the method and offer insights for the productive use of acceptability judgments with heritage language speakers.

Keywords. heritage language; acceptability judgments; methodology; experiment design

1. Introduction. This paper discusses the use of acceptability judgments in heritage language studies, and specifically with speakers of moribund heritage languages in the US. Acceptability judgments, previously also called grammaticality judgments, have been widely used in both descriptive and theoretical linguistics (see e.g., Schütze 2016; Schütze & Sprouse 2013). They are important – if not crucial – in cases where one wants to establish the boundaries of a linguistic system, i.e., which forms or sentences are *not* possible or acceptable in a language variety. Such negative evidence cannot be deducted from production data alone, as absence of a construction in a corpus is not evidence for its ungrammaticality. The structure may simply be highly infrequent or restricted to specific registers, dialects, or speaker groups.

Judgment data is also often used in heritage language (HL) studies. However, it has been claimed that acceptability judgments are not a suitable method for HL speakers (Orfitelli & Polinsky 2017). This controversy underlines the importance of discussing the method. This paper aims to contribute to this discussion by focusing on a specific group of HL speakers: the elderly speakers of moribund (Germanic) varieties in the US (cf. Putnam et al. 2018; D'Alessandro et al. 2022). §2 provides a brief introduction to acceptability judgments and discusses previous critique on this method. In §3 and §4, I present two studies that used judgment data from moribund HLs, specifically centering around the question of how this method can be used – or adapted – to gain insights into HL grammars. §5 summarizes, with the conclusion that acceptability judgments can be used, provided that certain (ethical and methodological) aspects are carefully considered.

2. Apparent paradox: AJTs in heritage language research. In an acceptability judgment task (AJT), a speaker is asked to judge whether a word or sentence is a "possible utterance of their language" (Schütze & Sprouse 2013:28). Judgments can be elicited informally or formally, and the judged sentences can be presented in written or spoken form. Various types of judgment tasks exist, ranging from forced-choice tasks (where participants choose which of two sentences is acceptable)

* Many thanks to the North American Norwegian participants of the study reported here. I am also grateful to the audience at WILA14, and two anonymous reviewers for their insightful feedback. Author: Yvonne van Baal, University of Stavanger (yvonne.w.vanbaal@uis.no).

 Bergen Language and Linguistics Studies 15(1), 7–14.
DOI: 10.15845/bells.v15i1.4548.

to yes-no tasks (where sentences are judged as either acceptable or unacceptable) to Likert-scale tasks (where sentences are judged on a numerical scale, e.g., 1-5). See Schütze & Sprouse (2013) for more discussion of the experimental design of AJTs. As mentioned above, judgments are a vital type of data for linguistics as they can provide information on what is *not* possible in a given language.

Orfitelli & Polinsky (2017) argue that AJTs are not a suitable method for HL speakers and should be avoided.[1] Despite this strong claim, AJTs are widely used in the HL field. For heritage Spanish in the US, for instance, Montrul & Ionin (2012) used a combination of experimental methods including an AJT to study the use and interpretation of articles. Another example is the study by Scontras et al. (2018) on number and gender agreement in nominal phrases, which was exclusively based on judgment data. The latter example shows that the method is used even by researchers who are critical to it. In other words, there is an (apparent) paradox: AJTs are widely used in the field, while they are also heavily criticized. Orfitelli & Polinsky (2017:198) raise "certain red flags about the validity" of AJTs and mention three problems in response patterns of HL speakers. I present these below, including some counter arguments to their critique.

The first issue raised by Orfitelli & Polinsky (2017) is the inconsistency in judgments, both between and within heritage speakers. At the same time, though, we know that there is some variation in monolingual speakers' judgments (Schütze & Sprouse 2013:45–47). Judgments are a type of behavioral data rather than a direct reflection of a speaker's grammar, and other factors may therefore influence judgments. In my view, there is no reason to dismiss data as invalid for the sole reason that they are not fully consistent. Rather, we have to be aware of inconsistencies and potential explanations for them.

The second problem that Orfitelli & Polinsky (2017) raise is that AJTs can provide different results than other comprehension methods, such as truth-value judgment tasks or picture-sentence matching. However, not all 'comprehension' methods are the same: AJTs test the grammatical *acceptability* of a sentence, while other comprehension methods measure its *interpretation*. For example, Montrul & Ionin (2012) used truth-value judgments to examine generic versus specific interpretations of Spanish plurals. Pure acceptability – in cases where negative evidence is required to determine which structures are impossible in the language – can typically not be tested other than with AJTs.

Finally, Orfitelli & Polinsky (2017) describe how heritage speakers tend to accept sentences that monolingual speakers reject, also known as the "yes-bias" (Polinsky 2018:69). In my view, this is mainly a problem when heritage speakers are compared directly to a control group of monolinguals doing the same task.[2] When comparing different conditions within the HL speaker group, differences in the heritage speakers' judgments are found (e.g., Scontras et al. 2018; Hopp & Putnam 2015). In these studies, the speakers differentiated between acceptable and unacceptable sentences, even when they did not reject the latter as strongly as monolingual speakers would.

Because of the mentioned issues, Orfitelli & Polinsky (2017) conclude that AJTs should only be used when there are no other methods available, and that AJTs (if used) should be part of a larger set of experiments. This seems somewhat contradictory, and as discussed above, I do not consider all the raised issues as problematic. However, they give us good reason to be careful in

[1] Orfitelli & Polinsky (2017) argue against the use of AJTs with "non-native speakers", by which they mean heritage speakers and L2-learners. Following, among many others, Rothman & Treffers-Daller (2014), I consider heritage speakers native speakers. I will not discuss L2-speakers in this paper.

[2] There are several other reasons to avoid comparisons with monolinguals, which I will not discuss here.

using this method. If the question of a study is whether something is possible (grammatical) in the HL, AJTs are the only viable method and combining with other comprehension methods is not necessarily insightful. However, I fully agree with the principle of triangulation and combining AJTs with production data of various types (e.g., conversations, storytelling, elicited production).

In addition to the challenges raised by Orfitelli & Polinsky (2017), there are specific challenges when working with certain groups of HL speakers. Many HL speakers of Germanic languages in the US are elderly, and they are the final generation of speakers (cf. Putnam et al. 2018). Factors that should be considered are that many of these speakers lack literacy in the HL and might be unfamiliar with the (present-day) standard variety. In addition, they may suffer from hearing difficulty related to their age. Some speakers are less used to following instructed tasks, and others may be uneasy with tasks that feel like a test. All these factors require specific adjustments in task design, but they may recall for conflicting adjustments: the lack of literacy can be circumvented with oral tasks, while hearing problems make oral tasks difficult.

Another challenge when working with elderly speakers of moribund HLs, is that there typically are few speakers in the communities. This means that we often work with (relatively) small groups of participants (cf. D'Alessandro et al. 2022). Performing statistical analyses may therefore be complicated, but that does not mean AJTs cannot be conducted at all with these speakers.

To demonstrate that AJTs can be used with speakers of moribund HL, and to illustrate the challenges and findings of such studies, I discuss two case studies in the next sections. Hopp & Putnam (2015) conducted an AJT with a group of Moundridge Schweitzer German speakers, discussed in §3. §4 presents the judgment data from North American Norwegian in van Baal (2020). For both studies, I focus on the methodology and on how the method was used to describe the languages; I do not discuss all findings in detail.

3. Case 1: Moundridge Schweitzer German. Hopp & Putnam (2015) investigate word order variation in Moundridge Schweitzer German (MSG), which is a moribund speech enclave in South Central Kansas. As is well-known, German is an SOV-language with V2 in main clauses but not in subordinate clauses. However, there are some restricted pragmatic contexts where V2 is licensed in the latter (for details, see Hopp & Putnam 2015:185–187). The aim of Hopp & Putnam's study was to investigate whether this asymmetric word order of German is retained in MSG.

They used a combination of production data and an AJT with oral stimuli. They collected data from 8 participants with a mean age of 85.2 years. The AJT consisted of 48 sentences (each preceded by a contextualization) that were recorded in Standard German and judged on a 6-point Likert-scale. Important here is that the stimuli were in Standard German, although the participants speak MSG and judged the acceptability of word order patterns in MSG. The speakers are to some extent familiar with Standard German. Hopp & Putnam (2015:198, fn.6) describe that the MSG-speakers only had minor difficulties in understanding the stimuli. These were mitigated by allowing participants to ask questions and hear the stimuli multiple times. The task took between 30-40 minutes per participant.

The results from the AJT show clear and statistically significant differences between certain syntactic conditions. In other words, even if a condition was not rejected categorically, there were clearly conditions that were rated low. The lowest rating was given to sentences in the 'word salad' condition,[3] and the other differences show that MSG "maintains the asymmetric German verb-

[3] This condition contained sentences with "an illicit word order that would nevertheless be semantically interpretable" (Hopp & Putnam 2015:199), which are ungrammatical by any standard.

second (V2) and verb-final (V-final) word-ordering closely tied to specific pragmatic information associated with clause-types and complementizers" (Hopp & Putnam 2015:180). Under certain conditions, V2 was accepted in subordinate clauses by the MSG-speakers, following systematic constraints from German syntax.

Hopp & Putnam (2015) present interesting results about word order patterns and syntax in moribund HLs. From a methodological perspective – my focus here – the study shows that AJTs are possible with this population and can provide insight into what is acceptable and unacceptable in the HL. The risk of the "yes-bias" (see above) was not problematic in this study that found clear differences between conditions and did not compare to monolingual speakers. The population in Hopp & Putnam (2015) is in many aspects quite similar to that of North American Norwegian, which is discussed next.

4. Case 2: North American Norwegian. Inspired by Hopp & Putnam (2015), van Baal (2020) conducted an AJT with a group of North American Norwegian speakers. North American Norwegian (NAmNo) is a moribund HL, spoken in the Upper Midwest of the US by elderly descendants of Norwegian immigrants prior to the 1920s. The study investigated double definiteness and word order (VO versus OV).

In Norwegian, definite phrases modified by an adjective require definiteness marking in both a prenominal determiner and a definite suffix, as in *den grønne bil-en* (DEF green car-DEF), 'the green car'. This is known as double definiteness. In production data, the prenominal determiner is often omitted, while the definite suffix is more stable (van Baal 2020, 2024). The aim of the AJT was to support these production data with acceptability data, as other types of comprehension data would not provide insights into the acceptability of phrases without the determiner (e.g., *grønne bil-en* 'green car-DEF').

The task consisted of two parts of 30 short sentences each, with a break between the two parts.[4] However, this turned out to be too long for the participants and they therefore completed half of the task. The different conditions were balanced across the two halves, and all participants judged sentences from all conditions. It was counter-balanced with which half the participant started, so all sentences have been judged by some participants. The oral stimuli were not spoken in standard Norwegian, but adapted to NAmNo. This means that they were spoken in an Eastern Norwegian valley dialect close to NAmNo, and contained specific NAmNo lexical items (e.g., *farm* rather than homeland Norwegian *gård* 'farm'). The stimuli were judged on a three-point scale that also had a "don't know"-option. The participants heard the sentence, could ask to hear it again and could ask for clarification. The task included an element of elicited imitation: participants repeated the sentence before they judged it. The AJT was completed by 7 elderly participants who had also provided production data earlier. They used between 18 and 25 minutes for (half of) the task.

The results seem less clear and optimal than those from Hopp & Putnam (2015) reported above. In general, the NAmNo speakers disliked the task and found it difficult. The task was difficult because of its length, which was mitigated by presenting only half of the task to each participant. The original task turned out to include a too ambitious number of items and conditions. Another complicating factor was that the participants struggled with hearing and understanding the stimuli, even though the sentences were adjusted to NAmNo. It seems that they had more difficulties with understanding the stimuli than the MSG speakers as reported in Hopp & Putnam (2015). The

[4] Example of a sentence with double definiteness: *Mannen liker den nye bilen* 'The man likes the new car'. Example of a sentence without double definiteness (no determiner): *Jenta ser hvite hesten* 'The girl sees the white horse'.

MSG speakers were somewhat familiar with the Standard German they heard, but I had expected that hearing NAmNo would be easier for the NAmNo speakers. One probability is that the MSG speakers – also being exposed to Standard German to a certain extent – are more used to variation in German, while the NAmNo speakers typically are not familiar with the written language (Hjelde 2015) and only with their own local dialect. Even if the difference between the two groups cannot be explained completely, the comparison shows that it is important to adapt tasks to each specific community.[5]

The participants responded orally to the stimuli and there is great variation in how certain they are in giving their judgments. Sometimes there are long pauses, or comments such as "I think maybe..." around the judgment, which express uncertainty. This is a complicating factor for the researcher, as it makes the judgments harder to interpret. At the same time, it is difficult to quantify the uncertainty and draw conclusions from it.

The task provides two useful (methodological) insights. The first insight is that the repetition data were very useful. The participants in van Baal (2020) repeated each sentence before they judged it. This was done to ensure that the speaker had heard the sentence correctly, but also because elicited imitation can provide indirect judgments. Here, the logic is that speakers who repeat a sentence do so with their own grammar, and a change in the sentence could thus reflect a correction of some sort (Vinther 2002).[6] This was indeed found in the data.

One relatively frequent case of repetition is the omission of the determiner that was present in the stimulus, as in (1). In these instances, the repetition changes the stimulus to something unacceptable in the baseline. This provides an indirect judgment that the omission of the determiner is acceptable for these speakers, in line with what the production data suggest. Furthermore, the repetitions of the AJT stimulus sentences provide examples where the participant changes a sentence to be more like the baseline. In these cases, we observe the addition of the definite suffix, as in (2). Despite the small data set and the challenges, the indirect judgments provide support for the production data collected earlier: they also show that the determiner is vulnerable for omission, while the definite suffix is more stable.

(1) a. Det store huset er veldig gammelt. (stimulus)
b. __ Store huset er veldig gammelt. (repetition)
'The large house is very old.'

(2) a. Jeg ser den svarte fugl. (stimulus)
b. Jeg ser den svarte fugl-**en**. (repetition)
'I see the black bird.'

The second insight that the AJT in NAmNo provides, is that there is variation across phenomena. The difficulties described above mainly applied to the sentences testing double definiteness in various forms. The (filler) stimuli testing word order (specifically, OV- versus VO-order) were easier for the participants and received very clear judgments. The ungrammatical OV-sentences

[5] A difference between the studies is that the stimuli were recorded by a woman in van Baal (2020), but by a man in Hopp & Putnam (2015). Female voices are typically higher, and higher tones are generally harder to perceive in age-related hearing loss (thanks to Joshua Bousquette for pointing this out).

[6] Cases of changes (rather than exact repetitions) are taken to be particularly insightful. When a participant changes the sentence instead of repeating it verbatim, this is taken as an indirect judgment from the participant on the original stimulus.

were rejected at a high rate, in fact, at the highest rate of all conditions. In all cases where an OV-sentence was judged acceptable, the speaker had changed or corrected it during the repetition.[7] This is also the only condition where speakers commented explicitly on what was wrong with it (typically with a comment about the sentence being "backwards") and participants furthermore judged these sentences with seemingly more confidence.

Interestingly, Hopp & Putnam (2015) also observe an asymmetry between different types of conditions. As described above, the MSG-speakers display a sensitivity to word order violations, but it is also found that they "did not show sensitivity in judgments to case-marker or subject-verb agreement violations" (ibid: 203), which were tested in separate conditions. Together with the data from NAmNo (van Baal 2020), this may suggest that AJTs work better for some domains. Specifically, they seem to be better suitable to elicit data on word order patterns than on agreement or clause-internal morphology.

5. Concluding remarks. There have been concerns about the use of AJTs with HL speakers, and the two studies discussed here illustrate that it can be challenging to conduct AJTs with elderly speakers of moribund HLs. However, both cases also show that it is not impossible: both Hopp & Putnam (2015) and van Baal (2020) obtained relevant insights into Moundridge Schweitzer German and North American Norwegian, respectively. Crucial in both studies was that the results from the AJT were triangulated with (elicited or semi-spontaneous) production data. This supports the recommendation from Orfitelli & Polinsky (2017) that AJTs, when used, should be combined with other types of tasks.

The studies on MSG and NAmNo compared different conditions within the HL rather than comparing the HL judgments with monolingual judgments. In both cases (although clearer in Hopp & Putnam (2015)), the heritage speakers made a difference between conditions. This type of comparison therefore makes the risk of the yes-bias less grave. The data in van Baal (2020) suggest that an element of elicited imitation, through repetition of the stimuli sentences, can provide useful insights into the given judgments. This is in line with Polinsky (2018:96) who notes that more implicit tasks may be preferred for HLs.

Both studies discussed here observed differences between linguistic phenomena that were investigated, in that the data on word order were much clearer and (in Van Baal's case) easier to elicit than data on agreement and phrase internal morphology. Future studies would have to investigate whether this asymmetry also is found in other HL groups. So far, the case studies reported here indicate that AJTs on word order can be conducted with HL speakers, even with the elderly speakers of moribund varieties.

At the same time, the challenges of AJTs with these speakers should not be underestimated. The task is generally difficult for the speakers, and it is crucial to adapt the task to the speaker population. This includes presenting stimuli orally. However, it may be hard to predict exactly which adaptations are necessary. While the oral stimuli in van Baal (2020) were adapted to the NAmNo variety, the speakers seemed to have more difficulty with the task than the MSG-speakers in Hopp & Putnam (2015). In general, it is important that participants can hear stimuli several

[7] The sentences in this condition contained a verbal complex, where the finite verb was in the correct V2-position, while the non-finite verb was placed before or after the direct object. Repetitions were analyzed as corrections when the sentence was changed to baseline-like VO-order, and as 'changes' when the response did not include a verbal complex. All responses in the latter category had baseline-like word order with V2 (see details in van Baal 2020:133–134).

times and ask questions about lexical items.

There is furthermore an important ethical dimension to this discussion. It is imperative to maintain a good relationship with research participants. Asking them to do a task that is very difficult or uncomfortable for them would be unethical. Most NAmNo speakers did not really enjoy the AJT, in part because the task was very long. This was 'solved' by shortening the task (see §4), but the participants all enjoyed story-telling tasks and semi-spontaneous conversations much more. For many research questions, AJTs may not be necessary, and then do not need to be conducted either to avoid the risk of unpleasant (and at worst, unethical) data collection.

Based on the experiences discussed in this paper, I would recommend any researcher planning to conduct an AJT with (elderly) heritage speakers (i) to keep the task short with a restricted number of conditions, (ii) to adapt to the population of speakers, and (iii) to combine it with other types of tasks. The latter does not only provide more varied data to allow for triangulation, but also makes the data collection sessions more enjoyable for the HL speakers, which is important from an ethical perspective. Adapting to the population (point (ii)) includes – but is not limited to – using oral stimuli adapted to the relevant variety or dialect, allowing participants time to get used to the procedure, giving them the opportunity to ask questions about stimuli, and maintaining a pleasant atmosphere. For all these aspects, it is important that the researcher knows the speaker population well and already established a connection with it.

The studies presented in this paper investigate moribund heritage languages, of which there are only few elderly speakers left who can provide data on their heritage language. In describing these languages before they disappear, we as researchers may need negative evidence collected through acceptability judgments. My conclusion is that acceptability judgments can be collected with this group of speakers and provide insights into heritage grammars, provided that ethical and methodological aspects are carefully considered in the design of the task as well as during data collection and analysis.

References

van Baal, Yvonne. 2020. *Compositional definiteness in American heritage Norwegian*: University of Oslo dissertation.

van Baal, Yvonne. 2024. Definiteness marking in American Norwegian: a unique pattern among the Scandinavian languages. *Journal of Comparative Germanic Linguistics* 27(1). https://doi.org/10.1007/s10828-023-09149-z.

D'Alessandro, Roberta, David Natvig & Michael T. Putnam. 2022. Addressing challenges in formal research on moribund heritage languages: A path forward. *Frontiers in Psychology* 12. https://doi.org/10.3389/fpsyg.2021.700126.

Hjelde, Arnstein. 2015. Changes in a Norwegian dialect in America. In Janne Bondi Johannessen & Joseph Salmons (eds.), *Germanic Heritage Languages in North America: Acquisition, attrition and change*, 283–298. Amsterdam/Philadelphia: John Benjamins Publishing.

Hopp, Holger & Michael T. Putnam. 2015. Syntactic restructuring in heritage grammars: Word order variation in Moundridge Schweitzer German. *Linguistic Approaches to Bilingualism* 5(2). 180–214. https://doi.org/10.1075/lab.5.2.02hop.

Montrul, Silvina & Tania Ionin. 2012. Dominant language transfer in Spanish heritage speakers and second language learners in the interpretation of definite articles. *The Modern Language Journal* 86(1). 70–94.

Orfitelli, Robyn & Maria Polinsky. 2017. When performance masquerades as comprehension: Grammaticality judgments in experiments with non-native speakers. In Mikhail Kopotev, Olga Lyashevskaya & Arto Mustajoki (eds.), *Quantitative Approaches to the Russian Language*, 197–214. London: Routledge.

Polinsky, Maria. 2018. *Heritage languages and their speakers*. Cambridge: Cambridge University Press.

Putnam, Michael T., Tanja Kupisch & Diego Pacual y Cabo. 2018. Different situations, similar outcomes. Heritage grammars across the lifespan. In David Miller, Fatih Bayram, Jason Rothman & Ludovica Serratrice (eds.), *Bilingual Cognition and Language. The state of the science across its subfields*, 251–279. Amsterdam/Philadelphia: John Benjamins Publishing.

Rothman, Jason & Jeanine Treffers-Daller. 2014. A prolegomenon to the construct of the native speaker: heritage speaker bilinguals are natives too! *Applied Linguistics* 35(1). 93–98.

Schütze, Carson T. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.

Schütze, Carson T. & Jon Sprouse. 2013. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

Scontras, Gregory, Maria Polinsky & Zuzanna Fuchs. 2018. In support of representational economy: Agreement in heritage Spanish. *Glossa: a journal of general linguistics* 3(1). 1–29. https://doi.org/10.5334/gjgl.164.

Vinther, Thora. 2002. Elicited imitation: A brief overview. *International Journal of Applied Linguistics* 12(1). 54–73.