



ORIGINAL ARTICLE

Exploring individual variation in Turkish heritage speakers' complex linguistic productions: Evidence from discourse markers

Onur Özsoy^{1,2}  and Frederic Blum³ 

¹Humboldt University of Berlin, Berlin, Germany, ²Leibniz-Center General Linguistics (ZAS), Berlin, Germany and ³Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Corresponding author: Onur Özsoy; Email: oezsoy@leibniz-zas.de

(Received 1 June 2022; revised 7 March 2023; accepted 2 April 2023)

Abstract

Research on multilingual speakers is often compared to monolingual baselines which are commonly treated as if they were homogeneous across speakers. Despite recent research showing that this homogeneity does not hold, these practices reproduce native-speakerism and monolingualism. Heritage language research, which established itself in the past two decades, is no exemption. Focusing on three predefined linguistic groups, namely Turkish speakers which are framed as monolingual in Turkey as well as two heritage bilingually framed groups in Germany and the USA, we ask: (1) Do heritage speakers of Turkish produce more discourse and fluency markers (FMs) than monolingual speakers? (2) Are the groups homogeneous, or is there wide variation between speakers across groups? We focus on the variation between and within groups using Bayesian Linear Regression with a multilevel model for speakers and heritage groups. Our findings confirm that the use of discourse and FMs is largely defined through individual variation, and not through the belonging to a certain speaker group. By focusing on variation across groups rather than between groups, our study design supports the growing body of literature that questions common heritage language research practices of today and shows alternative paths to understanding heritage grammars.

Keywords: bilingualism; heritage speakers; discourse markers; individual variation; heritage Turkish

Introduction

At least since Grosjean (1989), linguists have addressed the problem that research on multilingual speakers is often compared to monolingual baselines which are commonly treated as if they were homogeneous across speakers. This was the start of an ongoing push to rethink the idea of the native speaker, and a lot of literature has discussed alternative approaches (e.g., Bayram et al., 2019; Rothman & Treffers-Daller, 2014). Fundamentally, much recent research shows that the assumed

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

homogeneity does not hold (Castro et al., 2022; May, 2020; Shadrova et al., 2021; Shin, 2022). More specifically, Shadrova et al. (2021) investigated narrative production data in a well-controlled corpus of German native speakers which is often used for comparisons with L2 learners. They find large heterogeneity across morphological and syntactic subclasses within the monolingual German group. Before we delve an investigation of the homogeneity and heterogeneity in different groups, we need to define these terms.¹ We understand homogeneous groups as having similar or identical traits and behaviors. On the other hand, heterogeneous groups show varying and different behavior and traits between the members of a group. For our purposes, this becomes relevant if we want to compare different groups. Essentially, we ask whether monolingual and bilingual heritage groups are homogeneous among themselves, so they can be meaningfully compared to each other, or, whether these groups are more heterogeneous among themselves than between each other. We suspect the latter statement to be true and study this by looking at speakers' production of discourse markers (DMs) across different groups of bilinguals and monolinguals. Additionally, we believe that most researchers in our field refer to speakers who are "functionally monolingual" when they use the term monolinguals (Grosjean, 1989). This indicates that speakers only use one language on an everyday basis, but they might also have been exposed to other languages throughout their lifetime in school and with multilingual peers. Essentially, this implies the notion that is presented in viewing differences between monolinguals and bilinguals as a spectrum instead of a binary (Wiese et al., 2022).

Previous research on heritage languages

Heritage language research established itself in the past two decades and initially mainly investigated Russian and Spanish in the USA (Montrul, 2016; Polinsky & Scontras, 2020b). Around the early 2000s, it was known that people who were born in the USA (or immigrated there at a very young age) and had learned a language other than English from their family and home environment were showing diverging patterns from their peers in the homeland. In order to be able to better understand these newly emerging patterns, the heritage speakers were often starkly contrasted with so-called monolingual "controls" and "baselines" (Polinsky & Kagan, 2007).

We want to emphasize that research into heritage languages carries a bias by design: participants are defined by sociolinguistic groups rather than empirical linguistic data, that is, heritage speakers are defined as such because of where and how they live rather than because of how they produce linguistic phenomena. In fact, there are plenty of linguistic domains where studies robustly find no major differences between these predefined groups (e.g., Azar et al., 2020; Kupisch et al., 2017; Nagy & Gadanidis, 2021). For example, Azar et al. (2020) investigated Turkish–Dutch bilinguals' use of referential expressions in narrative production. They found that the speakers overall used the language-specific referential expressions appropriately. Therefore, it seems questionable why so many studies that investigate language variation in the setting of language contact and specifically heritage languages still frame a dichotomy between heritage speakers and monolinguals. However, we emphasize that we conceptualize heritage speakers

as native speakers too in line with most recent research in the subfield of heritage language studies (e.g., Rothman & Treffers-Daller, 2014; Wiese et al., 2022). This is based, among many factors, on the understanding that nativeness refers to naturalistic language exposure from early childhood.

The methodology has been quite uniform in heritage language research: Group means are gathered and compared to each other (Montrul, 2018). With advancing statistical methods such as mixed models, we can move beyond this tradition. There is potentially considerable variation between individual heritage speakers and within assumed groups, which is largely unexplored. This is the specific starting point which introduces the questions and directions of this study. Broadly, we ask how the interpretation of the results might change if we highlight the relevance of individual language use alongside group-centered analyses. While we might still expect to see indicators of group-characteristic language use, for example, heritage speakers might produce more discourse marks on average, we will focus on the variation within and between those groups. Such a perspective allows us to understand and critically evaluate the contrast between heritage speakers and monolinguals.

To measure individual patterns in speakers' linguistic productions, we look at discourse and fluency marker (FM) use, which is often taken as an indicator of speakers' competence (Degand et al., 2019; Fillmore, 1979; House, 2013; Simmons-Mackie & Damico, 1996). DMs can cover a wide range of parts of speech such as adverbs, nouns, prepositions, and conjunctions. They also carry many functions, some of which we can highlight here. In fact, DMs are defined by their function rather than by their part of speech. A DM might separate clause boundaries, or it could have a deictic function in that it puts attention to a clause or phrase before or after it (Fraser, 2006). Certain DMs develop functions that are specific to a certain speaker group framed by age or gender such as the use of *like* by young female speakers in North American Englishes (Tagliamonte, 2005). DMs can also signal to other interlocutors that a thinking process is going on and that the interlocutors should hold for a moment. FMs are often less nuanced and more covert compared to other DMs. We subsume FMs in the category of DMs for the purposes of this study. Again, a diverse range of parts of speech can be classified as an FM, but our study takes a narrow scope and limits the class of FMs to all types of hesitations including verbalized pauses such as *em* or *mm*. What combines DMs and FMs is that their frequency of use is for many speakers negatively correlated with a perceived proficiency in the language and (self-)confidence of the speaker (Belz & Odebrecht, 2022). As a rule of thumb, speakers with a higher use of DMs and FMs are perceived as being less proficient (Schmid & Fägersten, 2010). Some studies, such as Niebuhr and Fischer (2019), even were able to define 2–3 FMs per minute as a threshold between “elegant” and “notable” speech (Reed, 2000). As such, speakers in societies with prevalent monolingualism and native-speakerism are very attentive regarding the use of DMs and FMs. In the scope of this work, we do not engage with the specific functions and semantic-pragmatic aspects of DMs, but we rather view them as a broad complex linguistic category that allows us to show linguistic tendencies in the production of monolingual and bilingual speech. What is nevertheless certain is that the attention to DMs and FMs in heritage language research is sparse so far. Beyond our quantitative and variationist

exploration, many more studies on the functions of DMs in heritage varieties are needed.

The notion of incomplete acquisition has been popularized in heritage language research since at least 20 years (see Montrul (2002)) and much longer in other areas of linguistics (see Poplack and Sankoff (1984)). In its most basic understanding, the term *incomplete acquisition* implies that there is a full (normative) language or grammar that a person can acquire, but that this process has not been successful and therefore “incomplete” (e.g., Polinsky (2006)). Right after the term was introduced to be widely used, scholars problematized this term and asked questions about the nature of a “full language” (Cabo & Rothman, 2012; Putnam & Sánchez, 2013). The main arguments that we follow are twofold. First, it is a conceptual question whether one refers to speakers of a language that naturally acquired the language and use it on a regular basis for successful communication as native speakers or not. We echo Wiese et al. (2022) and others who locate heritage speakers on a continuum of native speakers with their own complete grammars. As we discuss below, also referring to Rothman et al. (2022)’s contribution in this special issue, this perspective captures the heritage data better and allows to ask more satisfactory research questions about heritage language grammars. Second, our argument is bound to a perspective of individual grammars. We can ask which different grammatical means speakers can use to express the same or similar meanings. For example, in our own research on clause combining in majority monolingual and heritage bilingual Turkishes, we confirmed previous observations that majority speakers seem to prefer subordination over coordination whereas this pattern is reversed in heritage bilingual populations (Özsoy et al., 2022). Both subordination and coordination are means of combining clauses which are equally valid and are guided by factors such as register and modality among other factors. So, instead of concluding for a deficiency in the heritage grammars, we found alternative syntactic strategies to express a similar grammatical function. This fine-grained investigation allowed us to learn more about the relationship between these context variables and clause combining in Turkish in general as opposed to just drawing a dichotomy between heritage and monolingual speakers. It also shows that comparisons between different groups of speakers can be beneficial in certain contexts if the groups and their varieties are viewed in their own right, and monolinguals are not just presented as the comparative norm (Rothman et al., 2022).

A more recent study compares Chinese L1 Mandarin speakers’ DM use with American L2 Mandarin speakers’ productions and offers more comparable data regarding the phenomenon under investigation (Diao & Chen, 2021). They found consistent significant differences between the groups regarding the frequency of use and the effect of the position for each DM that was investigated. Diao and Chen (2021)’s findings also highlight the importance for the consideration of a position effect, that is, the sentence position affects how many DMs are used. Crucially, to capture the different complexity in the use of DMs, we investigate them with respect to three different utterance positions. We interpret utterance-initial DMs as cases of macro-planning, utterance-medial DMs as instances of micro-planning events, and utterance-final DMs functioning as end points to ongoing trains of thought (Degand & Van Bergen, 2018; Fraser, 1990). Besides these speech-planning-related classifications of DMs and FMs, we are aware that DMs and FMs can also be used to

achieve certain phonological and prosodic patterns of a language like Turkish (Shriberg (2001) call this the “Acoustics Claim” and contrast it with the “Ecology Claim” regarding planning). Another characteristic of DMs and FMs is that their use is sensitive to different communicative settings, which we label registers (Tocaimaza-Hatch, 2018). Different social factors like familiarity and hierarchy govern the frequency of use of DMs (Brizuela et al., 1999). Therefore, register is an aspect that we will also consider in this study.

We explicitly investigated all DMs in the corpus (133 unique types of DMs) as this is a largely quantitative study that is interested more in the overall use patterns and dynamics that affect the use of DMs in general. Nevertheless, we want to exemplify three of the most prominent Turkish DM from our data which are *yani* “I mean,” *işte* “you know,” and *şey* “uh.” These forms have received some attention in the literature based on their use in Turkish spoken in Turkey (Altıparmak, 2022; Furman & Özyürek, 2007; Yılmaz, 2004). Yılmaz (2004) highlights that all three forms have multiple functions: *şey* is an element that expresses the speaker’s mental effort of lexical or structural recall, *işte* marks specific information in an utterance, and *yani* has a range of functions from a clause-connecting element to a clarification particle. Furman and Özyürek (2007) characterize these three forms as interactional DMs and observe different acquisition patterns in children. *Şey* is acquired first since it has a narrower function than the other two DMs *yani* and *işte* which have a broader range of functions and might not be fully acquired even by the age of 9 years. In addition, Altıparmak (2022) finds that these DMs are more common in spontaneous speech compared to planned speech. They also describe that adults ascribe more functions to these DMs compared to children, which is in line with Furman and Özyürek (2007). While these studies show that DMs in Turkish have manifold functions that deserve further investigation, we merely focus on their broader use in speech production to estimate their distribution in different groups (two heritage groups and one monolingual homeland group) and individual speakers.

Early groundbreaking studies in the field of heritage language research mainly found differences, especially some kind of nonstandard acquisition of the language, which has been framed in several ways as “incomplete,” “divergent,” or “attrited.” However, we think that this might be given in part at least due to the pressure to develop and publish new findings which focus on a new idea (Vasishth et al., 2018). For those studies, the idea was that heritage speakers are prone to differ given the different acquisition scenarios. And for many phenomena, especially those on the interfaces of syntax and pragmatics, these differences seem to be persistent. Many other studies could not find differences or overemphasize marginal differences which are not meaningful statistically or might just be task-driven. A nonexhaustive list of recent studies that falls in this domain is Azar et al. (2020), Kupisch et al. (2017), Nagy and Gadanidis (2021), and Oikonomou et al. (2022). To exemplify, Nagy and Gadanidis (2021) reanalyzed two morphosyntactic phenomena (pro-drop and classifiers) in heritage and homeland Cantonese speakers where differences between these groups had previously been reported. They conducted a variationist sociolinguistic analysis and found that both groups were equally able to utter complex linguistic structures. Among both of those groups, there were speakers who produced less complex structures and those who produced more complex

structures, but this variance is beyond any group variable. Here and for our purposes, we loosely define complex in contrast to simplex and mean that a linguistic form serves multiple functions depending on several sentence- and context-level factors. This shows how data can be (re)analyzed beyond group differences to explore the underlying factors of structured variation.

Such an analysis deviates from previous studies which often postulated rigid comparisons between a bilingual heritage group and a majority monolingual “baseline” group. As Rothman et al. (2022) point out as part of their contribution to this special issue, taking monolinguals as a baseline represents a *comparative bias* or even a *comparative fallacy*.² Monolingualism is not the default in language acquisition, and adding bilinguals as a “treatment” does not do justice to all the layers and variety of variables that effect (bilingual) language acquisition. To a large extent, this bias also might be the origin of dichotomous and oversimplifying labels such as “different,” “incomplete,” “divergent,” and “attitred.” Overcoming such labels and comparisons allows us to investigate the underlying factors that influence the variable realization of linguistic phenomena in a novel theoretical light. In related fields of heritage language linguistics and also within our field, few such studies already exist, but we seem to be at the turning point into a new decade of research allowing new theoretical perspectives which enable us to better address truly linguistic research questions.

We think that another fruitful extension to accompany new theoretical perspectives is to turn to recent statistical developments at the same time. By applying methods from a Bayesian statistical framework to our data, we can make more meaningful interpretations of our models (Gelman et al., 2013; Kruschke, 2015). For example, we can explicitly report and interpret any uncertainty involved in our inferences (Vasishth & Gelman, 2021). In general, all posterior distributions can be interpreted and the interpretation does not hinge on arbitrary p -values. This makes it possible to avoid the pitfalls of frequentist models and allows for more flexible model fitting. Another recent addition to inferential statistics that is not specific to the Bayesian approach are varying intercepts and slopes, which are often called “random effects” in frequentist statistics. Varying effects avoid that individual participants skew the estimate of an effect into a certain direction and also account for the nonindependence of data points (Baayen et al., 2008; Winter & Grice, 2021). Applying such varying effects in a Bayesian framework makes it possible to directly interpret the parameter values in question for any individual in the data set.

Apart from this detailed and focused analysis of individual variation, we also decided to divide the group variable in our data into three based on the sociolinguistic and sociocultural situation of the speakers. Most importantly, we labeled this variable as Country, which is short for Country of Elicitation. This helps us to address several distinctions that previous discussion around community cohesiveness and vitality have evoked (Iefremenko et al., 2021; Yagmur, 2011). It is a distinction that is not based on mono- versus bilingualism, but it rather captures the everyday linguistic practices of these groups. While Turkish is the dominant and official majority language in Turkey, it is a minority language in Germany and the USA. Among the latter two contexts, there are again different practices. An estimated 3.5 million people of Turkish descentance live in Germany

(Schührer, 2018). Estimates of the Turkish population in the USA range from 300,000 to 1,000,000. While still limited overall, in Germany, Turkish heritage language education is available at some public schools (Schroeder & Küppers, 2016), but in the US access to formal Turkish education is limited to Saturday schools (Otcu, 2010). These factors lead to different practices of speaking and using Turkish in the community: Whereas in Germany Turkish is available in public life and media and overall being used more, in the USA, it is mostly limited to contexts where it is spoken as a home language. These different practices are reflected in our conceptualization of the analysis toward the data. By including two distinct heritage speaker groups, we also aim to achieve better generalizability regarding our findings.

To summarize, our study jumps on the bandwagon of a lot of recent heritage language research in that it adds a variationist quantitative investigation of a specific phenomenon. This is in line with an overall shift in our field which has started more than a decade ago and has become more mainstream in recent years. We still expect to find some group-level variation, but highlight that it is equally, if not more important, to adequately account for the individual speaker variation. The use of DMs will therefore be driven more by controlled variables like register and utterance position as well as the notable speaker variation. The role of the group factor, which is often the main object of focus in studies of heritage language bilingualism, will be questioned allowing room for a continuum of fluent (first) language proficiency. This continuum allows bilingual native speakers to stand on equal grounds as monolingual native speakers, as suggested by Wiese et al. (2022).

Research questions and hypotheses

Above, we have argued that we can contribute to an ongoing shift in heritage language linguistics and bilingualism broadly, by bringing in new empirical evidence from heritage Turkishes and using cutting-edge statistical approaches to understand the role of individual variation in working with supposedly different (monolingual and bilingual) groups. The phenomenon under investigation is DM use which has generally received less attention in our field than other more purely grammatical phenomena. Since DM use is naturally prone to variation, it demonstrates an ideal study ground for between- and within-group variation. Based on these assumptions, and focusing on three predefined linguistic groups, namely Turkish speakers which are framed as monolingual in Turkey as well as two heritage bilingually framed groups in Germany and the USA, we ask our research questions. In doing so, we seemingly utilize a conceptual binary between monolingual and bilingual speakers. However, on several levels, our study is trying to overcome the binary. First, we do not use mono- versus bilingualism or even the group variable as a main independent variable in our model. Instead, it is incorporated in a nested random effect. This indicates that we do not view the speaker group status as a main driver of DM production. Second, the group variable that we utilize does not take one of the three speaker groups as a baseline. We sum-coded the variable which incorporates the idea that monolinguals cannot be an adequate baseline for bilinguals. Keeping these points in mind, our main research question first raises the ordinary binary that has been dominant in our field and then we proceed with the

second question to move beyond the binary, and we ask: 1) Do heritage speakers of Turkish produce more discourse and FMs than monolingual speakers? 2) Are the groups homogeneous, or is there wide variation between speakers across groups?

For our study, we conduct Bayesian Linear Regression to predict the probability for the use of DMs in different positions of the utterance. The model and hypotheses for the study presented in this paper were preregistered on OSF prior to analysis of the data.³ The authors of this study did not investigate the corpus regarding the predictions until the preregistration was submitted. In our preregistration, we postulate the following hypotheses:

Hypothesis 1. We predict that heritage speakers of Turkish in Germany and the USA produce more DMs than monolingual Turkish speakers overall. We distinguish between the use of DMs as macro-planning events in utterance-initial position, and as micro-planning events in utterance-medial position. Utterance-final DMs are expected to be fewer in numbers, as no planning is involved for the current utterance and DMs can be either due to abandoning a current utterance, or planning for the next one. We expect heritage speakers to show a larger use of DMs, and thus speech planning events, across all positions.

This hypothesis is motivated from the “monolingual baseline.” However, we do expect that in all groups, individual variation is more prevalent than group-level patterns. This leads to Hypothesis 2.

Hypothesis 2. With respect to the interpretation of our results on Hypothesis 1, we predict that the group-level effects can be overgeneralization because they allow us to make inferences about a whole group even though there might be large individual variation. In particular, we believe that the group effect is largely due to “influential individuals,” while most other heritage speakers will produce as many DMs as monolingual speakers. We hypothesize that the use of DMs is highly individual.

What we want to test with this hypothesis is whether all speakers in a group behave in a coherent way compared to the speakers of other groups, or if any resulting pattern is observable on the individual-level only. Those “influential individuals” could then influence the model in such a strong way that it appears as if there were a systematic pattern on the group level, even though there isn’t. For our analysis of DMs, this means that a large portion of them will produce as many discourse and FMs as monolingual speakers. We based this prediction on other studies that were able to attribute changes in heritage grammars to certain individuals in the groups (Goschler et al., 2020; Iefremenko et al., 2021; Özsoy et al., 2022). Our careful analysis of individual variation will be able to locate these speakers and adequately allocate between-group differences to those speakers. While recent advances in statistical modeling have started to address speaker variation with varying intercepts and slopes, our approach focalizes the differences highlighting that they should sometimes receive the main attention when between-group analyses are conducted, instead of merely controlling for individual variation and discussing only the group-level effects. That is especially important for

phenomena that naturally lean to large variation such as the production of DMs and FMs (Sankoff et al., 1997).

Further, we predicted the following results for the control variables:

- (a) Participants will produce more DMs in the informal register setting than in the formal register setting.
- (b) Larger utterance length will facilitate the use of DMs as lexical access becomes more demanding and more planning is involved.

Methods

Positionality statement

Both authors are at the beginning of their academic career. We believe that many concepts in linguistics, and in the study of heritage languages specifically, should be revised carefully. The first author is a young male person of color who was socialized in Germany's capital Berlin and acquired Turkish and German bilingually. He identifies as Turkish-German and is part of a minority community in Germany which has and still is facing marginalization and (structural) racism on many levels including language. This status affects his view toward heritage speakers as many of the people in his social networks are heritage speakers who use the heritage language vitally and vividly everyday. Therefore, speaking of heritage languages in a deficit-oriented way seems alienating to him. He has experienced denial of native speaker status in both of his native languages by authorities such as teachers and in work life by colleagues and in academic articles. These experiences make the author particularly sensitive to the practical implications that descriptions of heritage speakers might have to individuals' lives. The second author is a young White male, raised in an academic monolingual environment. His perspective on the research is mainly from outside point of view. Regarding this study which discusses heritage languages, we think that we benefit from the fact that one author belongs to the in-group of Turkish heritage speakers and the other belongs to the out-group. This led us to explain our own perspectives and opinions to each other and to reevaluate those perspectives together. In the process of writing this study, we questioned both perspectives continuously and developed a shared stance on the role of individual variation and its importance to linguistics.

We think that the Bayesian mixed models are a promising methodological approach to statistics in linguistics. Their flexibility in modeling data and, in our case, the possibilities of modeling variation and uncertainty are a huge argument in favor of leaving frequentist statistics behind. While both approaches share some deficiencies, the Bayesian approach makes it much easier to make all problems transparent.

For example, low sample sizes for some speakers make the estimation of parameter values difficult. This is also a problem in our study. There is quite some variation with respect to the amount of data each speaker contributes, which results in some large uncertainty intervals.

However, these uncertainties are made explicit in our reporting of the model fit. By fitting a Bayesian model with varying effects, we try to capture the effects both at group and at the speaker level.

We also want to include a “Constraint on Generality” (Simons et al., 2017). A main limitation of our study is that the perspective on DMs is purely quantitative. We do not explore the semantic role of specific DMs, nor do we investigate possible differences between the groups in the use of specific DMs. All our conclusions thus generalize only with respect to the quantitative use of DMs.

Overall, the results fit well with our hypotheses on the underestimated importance of individual variation in linguistics. Through the preregistration of our study, we tried to avoid any bias in the interpretation of our results as good as possible.

Data

In order to secure data sustainability and reproducibility and to harvest the synergistic quality of collaborative annotation, we decided to use the openly available corpus of the “Research Unit Emerging Grammars in Language Contact Situations: A Comparative Approach” (RUEG, Wiese et al., 2022). All code and the processed data data are curated on Github (<https://github.com/Tarotis/exploring-individual-variation-in-turkish-heritage-speakers-complex-linguistic-productions>) and published on Zenodo (<https://www.doi.org/10.5281/zenodo.7838068>).

Data from 188 participants were analyzed. It is made up of data from three countries, namely Germany ($n = 65$), Turkey ($n = 66$), and the USA ($n = 57$). Within each country, half of the participants were adults and the other half were adolescents. The sample size was determined based on statistical power considerations and feasibility limits for this large-scale cross-linguistic project by the original research group that built the corpus. Each participant gave narrations in two modes (spoken and written), two settings (formal and informal), and two languages, resulting in eight documents/narrations per participant. In our study, we only investigate the spoken Turkish data. The elicitation of the data was based on a narration task of a fictional event. A stimulus video presenting a nonsevere car accident at a parking lot was shown to every participant. The task was to imagine having witnessed the accident in person and to narrate what happened, both orally and in writing. Following the “language situations” setting (Wiese, 2020), two distinct communication situations (formal and informal) were elicited. The formal language situation simulates a communication with a police officer, while the informal setting is communication among peers. We coin the variable that includes the formal and informal settings as “Register.” We include both settings in our analysis and use this variable as a fixed-effect predictor in our model.

The Turkish-speaking participants from Turkey were raised monolingually and have not been exposed to another language until they started English lessons in primary school around the age of 10 years. Turkish is still the only language that these participants use regularly which allows us to describe them as “functionally monolingual” (Grosjean, 1989). Additionally, they are also not exposed to any other language except Turkish in everyday communicative settings with other people.

The speakers were all born and raised in the two major western Turkish cities, Eskişehir and Izmir, where mostly the standard variety of Turkish is spoken.

We further characterize the participants, by presenting general demographic and linguistic detail about them. Namely, these are age, linguistic proficiency, and Turkish language exposure. However, we do not regress these background variables in our model as we do not have any theoretical or conceptual reasons to assume that these would affect the use of DMs. We present these background variables by referring to the Country variable with its three levels: Germany, Turkey, and USA. Since age was controlled for in the elicitation of the data, it is similar across all groups (Germany $\bar{x} = 21.57$; Turkey $\bar{x} = 22$; USA $\bar{x} = 21.86$). For the bilinguals in Germany and the USA, we can say that the speakers are balanced on average based on their self-reported Turkish language exposure among their languages (Germany $\bar{x} = 52\%$, USA $\bar{x} = 53\%$). Additionally, the participants self-reported their proficiency across four linguistic domains (reading, writing, listening, and speaking) with a scale from 1 to 5 for each of these domains summing up to a maximum score of 20. Overall, the mean ratings indicate that our participants were (highly) proficient users of Turkish (Germany $\bar{x} = 16.64$; USA $\bar{x} = 15.14$).

The data were automatically tagged and manually annotated by Turkish-speaking reviewers who are not part of this current study. Therefore, they were unaware of the predictions in this study.⁴ The corpus in its current version (0.4.0) is openly available under a CC0 1.0 Universal license.⁵ We downloaded all the data of the corpus with POS tags and utterance ID available for each token and filtered the data for spoken modality. Due to our focus on spoken discourse, we do not investigate the written data that are also part of the available corpus.

The data we use for our study were exported from ANNIS and automatically annotated by ourselves for position, utterance length, and DMs. We will elaborate briefly on each of these. DMs are already tagged as such in the RUEG corpus with a dedicated part-of-speech tag and include FMs. We further decided to treat multiple DMs following each other within the same utterance as being a single constituent, so they are only counted once in our model. For example, the idiomatic expression *kolay gelsin* “May it come/feel easy to you” falls in this category. Utterance length was computed as the number of all tokens in the utterance that are not DMs. This count was centered and standardized. With respect to “Position,” all tokens, including DMs, were annotated for initial, medial, or final position within their utterance. This annotation process was realized fully within Python and is replicable through the published code.

Table 1 gives a first overview of the total number of tokens as well as the number of DMs across register, groups, and positions. It stands out that in utterance-final position, there are very few DMs. In absolute numbers, most DMs occur in utterance-medial position, but once taking the higher number of total tokens in that position into account, utterance-initial position has the highest relative amount of DMs with respect to the other positions. This is shown in Table 2.

Bayesian linear regression

We will use a Bayesian Linear Regression model for our study. The main reasons to use Bayesian models is the flexible fitting of models as well as the direct

Table 1. Total number of tokens and discourse markers per register, group, and position

Item	Group	Register	Initial	Medial	Final	Total
Overall tokens	Heritage speakers in Germany	Formal	793	4657	997	6447
		Informal	906	4260	1030	6196
	Monolinguals in Turkey	Formal	544	4067	704	5315
		Informal	781	3707	862	5350
	Heritage speakers in the USA	Formal	648	3650	816	5114
		Informal	702	3023	812	4537
Discourse marker	Heritage speakers in Germany	Formal	271	373	17	661
		Informal	252	348	78	678
	Monolinguals Turkish	Formal	226	234	13	473
		Informal	197	198	69	464
	Heritage speakers in the USA	Formal	208	290	11	509
		Informal	174	257	35	466

Table 2. Relative amount of discourse markers per position

Item	Initial	Medial	Final	total
Overall tokens	4374	23,364	5221	32,959
Discourse markers	1328	1700	223	3251
Relative amount of discourse markers	30%	7%	4%	10%

interpretability of the results, avoiding many of the frequentist pitfalls (Nicenboim & Vasishth, 2016, p. 3). Instead of *p*-values and significance levels, all parameters of the fitted model have a posterior distribution which is interpretable as containing the true values, given the prior distributions and the data (McElreath, 2020, p. 58). The posterior distribution can further be read as quantifying the uncertainty involved in the estimation of the true values for each parameter (Gelman et al., 2013, p. 11). The possibility to report uncertainty involved in statistical inference is not exclusive to Bayesian methods, but the direct interpretation of the posterior parameter distribution facilitates the explicit discussion of uncertainty that is necessary in order to account for any responsible assessment of hypotheses (Vasishth & Gelman, 2021). A common way of reporting on the posterior distribution is by the “Highest Posterior Density Intervals” (HPDI) (Kruschke, 2021, p. 1286). In order to avoid misinterpretation as frequentist confidence intervals, we will generally report the 89% HPDI (McElreath, 2020, p. 58), which has become the standard in some Bayesian applications (Makowski et al., 2019). This interval contains 89% of the true values of a parameter, given the prior and the data. In our plots, we also make use of the 99.7% HPDI, given that this value captures an important relation to the standard deviation of a normal

distribution. While the 95% interval corresponds to approximately all data points within two standard deviations of the mean, the 99.7% interval contains all data points within three standard deviations (Leys et al., 2013, p. 764). By this, we maintain a relation between the standard deviation of the distributions and the displayed parameter values.

Another advantage of Bayesian statistics is the incorporation of prior information in order to arrive at the posterior distributions (Gelman et al., 2013, p. 24). The priors are a crucial part for any model building and represent restrictions on possible parameter values. They are formulated as prior distributions of parameter values. Their most common application is as “weakly informative priors,” which limit the computational space of a model within boundaries that are designed in order not to influence on the posterior distribution but facilitate the computational process of modeling (Gelman et al., 2013, p. 51). For example, standard deviations can be modeled by an exponential distribution, which limits their values to positive numbers. Predictor variables can be modeled with a normal or Student’s t-distribution, which is centered on 0. This design excludes unreasonably large values of parameters but does not introduce bias with respect to the inferences.

A nontechnical version of *Bayes’ Theorem* underlying all Bayesian statistics is presented as Equation 1 (McElreath, 2020, p. 37):

$$\text{Posterior distribution} = \frac{\text{Probability of data} \times \text{Prior}}{\text{Marginal Likelihood}} \quad (1)$$

Results

Model

In order to evaluate the first hypothesis, the models will be compared both from a quantitative and a qualitative perspective. The null model does not assume any grouping of the speakers and only computes effects on the speaker level. The alternative hypothesis groups the speakers according to their sociolinguistic background as either being a monolingual speaker living in Turkey or having a bilingual background with German or US English, respectively.

The first model comparison is done via Bayes factors (BFs) and bridge sampling. The BF is the rate of the marginal likelihoods of a null and an alternative model. Bridge sampling is a methodology to compute the BF a certain amount of times to report the mean and standard deviation in order to report the stability of the BF over different sampling runs (Schad et al., 2022). The stability of the BF is important in order to avoid that a single BF computation is biased through the random sampling process involved. This workflow will show whether the grouping captures the variation in the data in a more efficient way than the null model. We will take $\text{BF} > 3$ as an arbitrary threshold for one of the models if the value is consistent after bridge sampling. The threshold of 3 has been established as a minimum value for providing some form of evidence for an alternative model (Jeffreys, 1939). However, the most important criteria for our study is not the computation of BF, but rather the manual assessment of differences in the posterior distributions between the

groups. The quantitative assessment of the models will only decide whether we use the grouped or the ungrouped model for further inferences.

Following the quantitative analysis, we compare the varying slopes between groups and/or speakers for all positions in order to see whether the observed patterns are coherent trends within each speaker group. Together, both analyses will be combined to evaluate Hypothesis 1. The second hypothesis will be evaluated qualitatively by comparing the variation between groups to the variation between individuals. The standard deviation between speakers and groups as well as the credibility intervals of the varying effects per speaker will be the most important evaluation metrics in this part of the study. Crucially, we predict that the individual variation is substantially higher than the variation between groups.

Our models calculate the probability of any item in an utterance being a DM, given a variety of predictors. The response distribution will be a Bernoulli distribution because we coded the target variable as binary. The model variables, annotated for each item in the utterance, are the following:

- DM (binary)
- Utterance length (numeric)
- Register (categorical: “Formal” and “Informal” elicitation setting of utterance)
- Position: (position of item in utterance: Initial, Medial, and Final)
- ID of speaker (factor with 188 levels, nested into three groups)
- Group (factor with three levels: monolingual, heritage speaker in the USA, and heritage speaker in Germany)

“Group” and “speaker” will each be modeled with varying effects with regard to the utterance position of the token (Baayen et al., 2008; Gries, 2015). This is necessary in order to account for the nonindependence between the data points (Winter & Grice, 2021, p. 1258). Further, we want to investigate closely whether the effect we can observe for any effect on the group level is actually due to an universal tendency or an artifact of highly influential individuals. This can be done by explicitly analyzing the varying slopes per speaker. Furthermore, in the grouped model, speakers will be modeled as nested within the factor “Group.” This means that we tell the model that within our data set, each speaker does always belong to one single group. In the ungrouped model, the “Group” variable will be dropped.

Utterance length and register will be used as fixed effects predictors. We code the target parameter “Position” with an index coding approach (McElreath, 2020, p. 156). This means that we do not calculate an overall intercept but rather compute the intercept for each of the three factor levels. This is not a common type of contrast coding (Brehm & Alday, 2022; Schad et al., 2022) but has advantages in our specific case due to the easier interpretation of all intercepts (Kurz, 2021). This causes the model to compute more parameters but facilitates their interpretation by establishing an intercept for each index. Given the large expected differences between the levels, this makes the interpretation of our results much easier, as the group- and speaker-specific slopes can be directly compared to the overall intercept for each position. We balance this additional computational load by more warm-up and sampling iterations and by including more informative priors than previously intended.

Table 3. All predictors with their prior distribution and short description

Parameter	Prior distribution	Short description
Position	Normal(0,2)	Initial, medial, and final Position of item in utterance
Utterance length	Normal(0,2)	Standardized number of nondiscourse marker tokens in utterance
Register	Normal(0,2)	Formal and informal register
Group	Exponential(10), LKJ(12)	Effects and correlation between parameters for the three groups heritage – Germany, heritage – US, and monolingual speakers
Speaker	Exponential(10), LKJ(12)	Speaker-level effects and correlation between parameters

The prior distributions for all parameters are given in Table 3. While a prior based on the normal distribution includes values within a range of -4 to 4, the Exponential distribution limits standard deviations to be positive. The Lewandowski–Kurowicka–Joe (LKJ) distribution models the correlation between varying intercepts and slopes and includes values between -1 and 1. All priors can be considered “weakly informative” and limit the parameter space by assigning a low probability to extreme values. This was done by running the same model by sampling only from the priors and comparing the results to the probability space.

The model is fitted with the formula as presented in Equation 2. We fit the models in brms (Bürkner, 2017), an R-interface (R Core Team, 2022) to STAN (Carpenter et al., 2017), a probabilistic programming language. The model included 4,000 warm-up iterations with a total of 20,000 Markov Chain Monte Carlo (MCMC) draws. Eight cores were run in parallel in order to have eight separate MCMC chains. The “adapt_delta()” value is at 0.98. The models had a fixed seed at 42 in order to be reproducible. For preprocessing, we made use of various tools offered by the R-package “tidyverse” (Wickham et al., 2019). The plots are made by a combination of “tidyverse” tools and the package “bayesplot” (Gabry et al., 2019):

$$0 + \text{Position} + \text{utterance length} + \text{Register} + (0 + \text{Position} | \text{Group} / \text{Speaker}) \quad (2)$$

All \hat{R} -values were below 1.01, which is a necessary but not a sufficient condition to confirm model convergence (Vehtari et al., 2021). We further used visual posterior predictive checks to show that the models correctly predict the data (Gabry et al., 2019). Figure 1 shows two different plots for the grouped model, showing that predictions fit well with the original data for each group.

Given that the quantitative and qualitative measures of model convergence both show that the models are fitted as intended, we can turn to the evaluation of the two models with respect to the hypotheses.

Hypothesis 1

In the first step, we compare the null model to the alternative model using BF. Running 10 unseeded computations of both models, the average BF is 33.6 with a

Table 4. Model comparison using the expected log pointwise predictive density and PSIS-loo

Model	ELPD	Standard error of ELPD	PSIS-loo	Standard error of PSIS-loo
Grouped model	0.00	0.00	19,745.54	238.66
Ungrouped model	-3.41	1.90	19,752.37	238.75

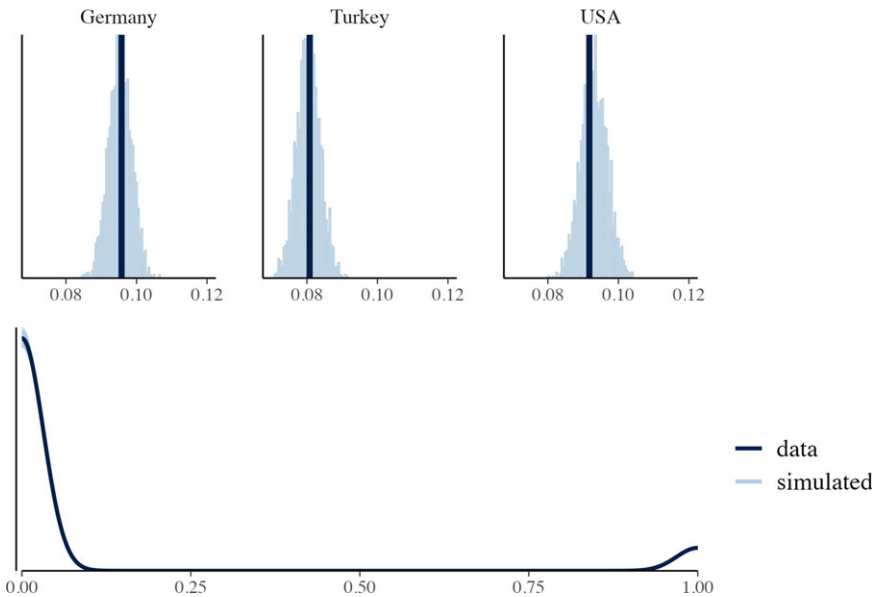


Figure 1. Posterior predictive checks for the grouped model across all three groups of heritage speakers.

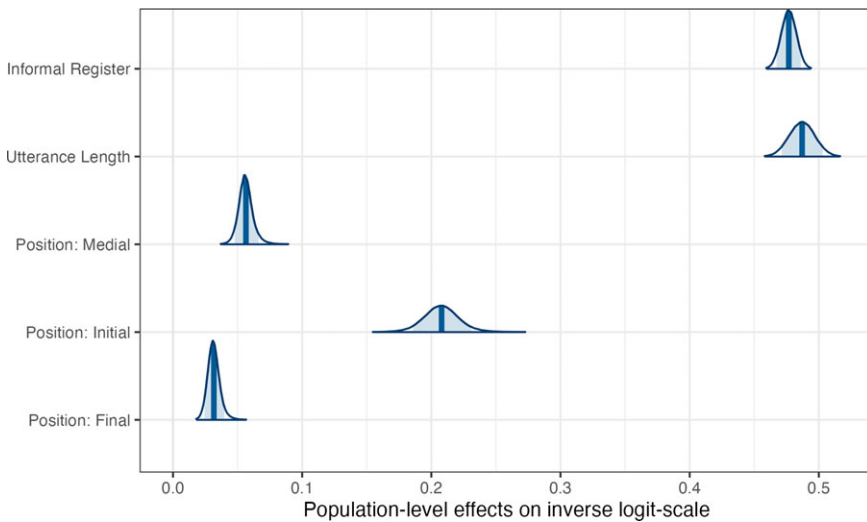
standard deviation of 15.9. This can be considered weak evidence in favor of the grouped model. However, the standard deviation is relatively large, which shows that the BF computation is not stable. In order to further investigate whether this variation is meaningful in any way, we also compared the models using information criteria that describe their predictive performance (Vehtari et al., 2017). We use the expected log pointwise predictive density for this, which is calculated through Pareto-smoothed importance sampling for leave-one-out cross-validation (PSIS-loo, which is reported in Table 4).

The expected log pointwise predictive density (ELPD) is the key value for interpretation and stands for the theoretical expected log pointwise predictive density. The best model in the comparison is set at 0, and all other models are compared to this best model. In our case, however, we see that both for the ELPD and the raw leave-one-out cross-validation (LOO) value, both models are very close to each other. Crucially, both measures are within two times the standard error of each other.

Table 5 presents the results for the population-level variables for the grouped model. They are further visualized in Figure 2. The thick blue line in the center of each parameter is the estimated mean of the posterior distribution. The light blue area corresponds to the 89% HPDI, whereas the total distribution reflects the 99.7%

Table 5. Population-level effects for all model parameters on the inverse-logit scale

Parameter	Estimate	89% HPDI
Final	0.03	0.02 to 0.04
Initial	0.21	0.19 to 0.23
Medial	0.06	0.05 to 0.07
Informal register	0.49	0.47 to 0.50
Length of utterance	0.48	0.47 to 0.49

**Figure 2.** Effect of population-level predictors on inverse-logit scale.

HPDI. Due to the outcome distribution chosen, the raw model output is on the logit scale, but for better interpretation, all results are presented on the inverted logit scale. This means that all values can only be interpreted as probability in isolation of the other parameters, which are held constant. The numbers reflect the probability of any item being a DM.

Confirming our predictions, both informal register and utterance length have a strong impact on the occurrence of DMs. Both HPDI are between 0.45 and 0.5, nearly doubling the probability in the informal register, or each unit of standardized utterance length, respectively. This means that both in the informal register and in longer utterances, DMs are much more likely.

The results with respect to the different speaker groups and utterance position can be summarized as follows:

1. For all groups, the probability for a linguistic token to be a DM is highest in utterance-initial position (~21%). In utterance-medial position (~6%), DMs are more likely than utterance-finally (~3%), but the difference to the initial position is substantial.

Table 6. Probability for discourse markers across groups and positions

Parameter	Group	Estimate	89% HPDI
Initial	Germany	0.21	0.17 to 0.25
	Turkey	0.22	0.18 to 0.27
	USA	0.20	0.16 to 0.24
Medial	Germany	0.06	0.04 to 0.09
	Turkey	0.05	0.03 to 0.07
	USA	0.06	0.04 to 0.08
Final	Germany	0.03	0.02 to 0.05
	Turkey	0.04	0.02 to 0.06
	USA	0.03	0.01 to 0.04

2. In all positions, the HPDI for all the groups largely overlap, indicating that there is no categorical difference between the groups. All differences are small tendencies, and the large uncertainty intervals indicate additional sources of variation within the “Group”-variable, that is, large individual differences.
3. In utterance-medial position, speakers from the “Turkey” group are less likely to produce DMs.
4. In utterance-initially and utterance-finally, speakers from the “Turkey” group are slightly more likely to produce DMs.

With respect to position, the most likely position for a linguistic token to be a DM is utterance-initially. This resembles the general distribution of tokens and DMs as in Table 1. Further, it is still twice as probable to have a DM in any medial than in final utterance position, even though the difference is rather small. This confirms our prediction with respect to the low number of utterance-final DMs. As there is a substantial number of utterances that only consist of one or two DMs and no lexical elements ($n \approx 200$), we run a model filtering out those data points in order to ensure that the results were not inflated by those cases. Contrary to what we feared might happen, all estimates were stable, and changes were below 2% on the population level in all positions.

We predicted that heritage speaker groups are more likely to produce discourse and FMs across all positions. The values for each group are presented in Table 6, and Figure 3 compares the effects of all groups across the different utterance positions visually. For all groups, the likelihood of featuring utterance-initial DMs is higher than in the other two positions. In utterance-final position, DMs are least likely, and the numbers are close to zero. With respect to the between-group comparison, the results differ from position to position. For the “Initial” position, the effects largely overlap. There is a very weak tendency for speakers from the “Turkey” group to feature more DMs in this position than the other two groups, but the difference is quite small and the HPDI are still largely overlapping. A slightly larger difference occurs in utterance-medial position, where the “Turkey” group shows an overall lower probability of featuring DMs than the other two groups. In utterance-final

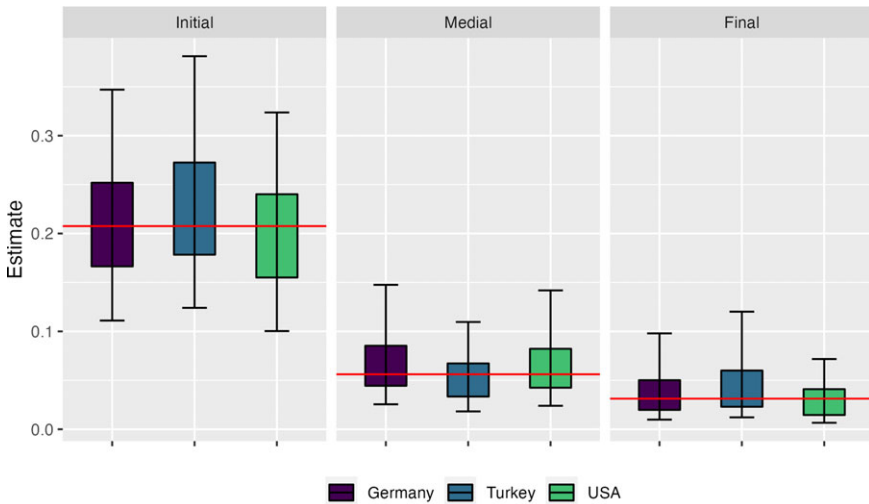


Figure 3. Group-level effects for the probability of discourse markers per utterance position, with intercept per position (red line), 89% HPDI (boxplot), and 99.7% HPDI (error bar).

position, the results are similar to the initial position: the effects largely overlap but the “Turkey” group shows an overall slightly larger effect.

Across all levels, the HPDI largely overlap, which shows that there are no strong differences between the three groups. Further, the comparison of the predictive information criterion shows that there are basically no differences with respect to predictive power between the grouped and the ungrouped model. The lack of differences between all groups is highlighted through the size of the HPDI of the parameters, which are about the same size for all groups. This means that the variation within all groups is roughly the same.

With respect to effects in different positions, there are some small differences, but no absolute patterns. In utterance-medial position, there seem to be some small differences between the monolingual group and both heritage groups, but even there the posterior intervals are largely overlapping. As an interim conclusion, we can say that the model does not pick up any conclusive differences between monolingual and heritage speakers. It does show, however, that the presence of DMs in an utterance depends largely on the position of the item in an utterance. DMs are far more frequent in utterance-initial than in utterance-final or -medial position.

Hypothesis 2

For Hypothesis 2, we want to investigate whether the results are coherent for all speakers within each group. In Table 7, we present the standard deviations for the estimates of all speakers for each speaker group, as well as the standard deviation for the group-level estimates. The table shows that the level of variation is much higher within all three speaker groups than at the group-level variable. This is further reflected by the standard deviations on the logit scale for both the Group (0.17) and Speaker (0.64) parameters. The much larger standard deviation for the Speaker

Table 7. Standard deviation of estimates on probability scale within speaker groups and the group-level parameter

Group	Final	Initial	Medial
Speaker: Germany	0.09	0.12	0.13
Speaker: Turkey	0.08	0.11	0.11
Speaker: USA	0.08	0.14	0.14
Group level	0.009	0.006	0.005

variable shows that in all positions, there is more variation between speakers than between groups. This already hints at the important fact that there is a lot of variation in the data that the grouping variable cannot capture. Given the low overall amount of DMs in utterance-final position, it is theoretically possible to argue that, at least for this position, the effect is due to low sample size. However, this is not the case for the other two positions that feature many more data points. Thus, the most likely explanation is that the numbers are consistent across all three positions.

Another important observation is that there are no substantial differences of the standard deviations between the three groups. The only difference that perhaps could be of interest is between the monolingual speakers and the speakers of the US heritage group, which show a difference of 0.03 for the word-initial and word-medial level. Given the overall values for those positions, however, it is of doubt whether those small differences should be overinterpreted. This difference is much smaller than the standard deviations between individual speakers of any group, which, in consequence, emerges as a source of larger variation.

Visualizing the speaker-level effects is difficult, given that we are discussing data from 188 individuals. In a first overview, the speaker-level results are presented in Figure 4. The effects are split in a grid across groups and positions where all speakers are ordered according to their estimated effect. The red line represents the population-level mean for the respective position across all groups and speakers. If in any part of the grid there are white areas, that means that the respective speaker group does not have speakers in that area of probabilities. We have two main observations. First, few individuals from the two heritage groups are among the most likely to produce DMs in utterance-medial position. However, this is limited to a very small number of speakers and does not seem to be consistent for the whole group. Second, no speakers from the monolingual group are among the least likely to produce DMs in utterance-final position. This could be interpreted as a small statistical trend that monolingual Turkish speakers tend to produce more DMs in this position. However, at least for the utterance-final position, this interpretation might be misleading, because the probabilities are all quite close to zero, and the HPDI largely overlap.

We also adapted the plot from Figure 3 for all speakers. However, due to the large amount of speakers in the data, we decided to move these plots to the appendix. This makes it possible to compare the different slopes across all three positions for each speaker, instead of only comparing the position-specific slopes in each group. These plots further underline the variation between speakers within the same groups.

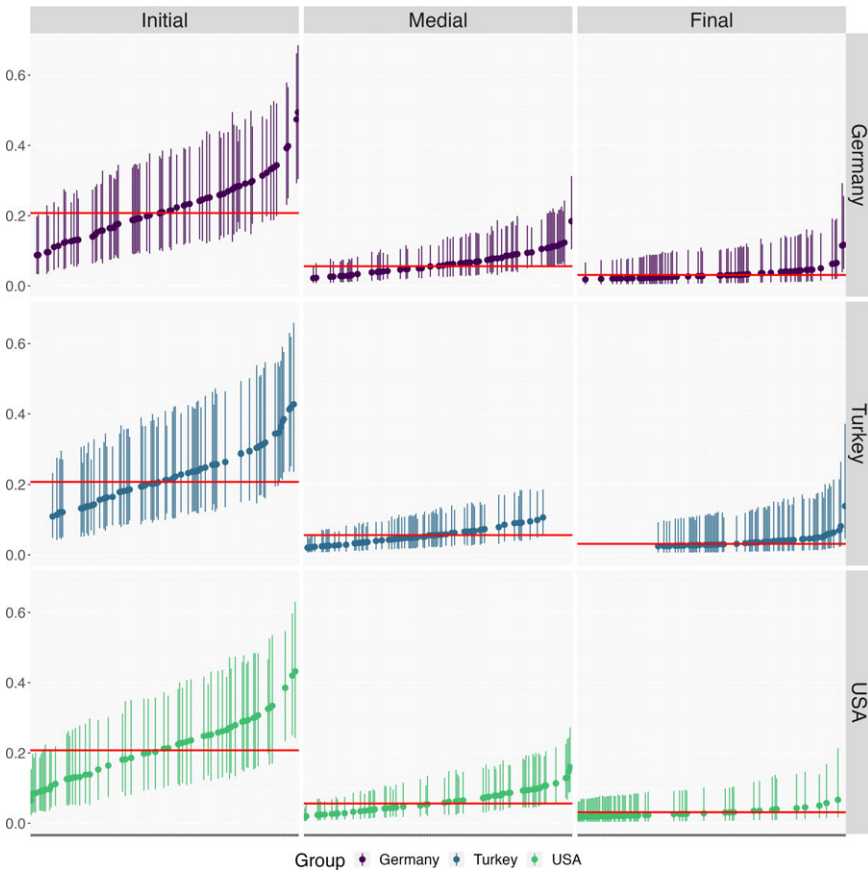


Figure 4. Speaker-level effects for the probability of discourse markers per utterance position.

This becomes clear by visually comparing the individual confidence intervals. Across groups, individuals vary especially with respect to the amount of word-medial and word-final discourse makers, while all individuals seem to produce more discourse makers in the initial position. However, they do so to different extents. While utterance-position seems important, it does not explain all the variation either. As already indicated by the standard deviations, there is a large amount of individual variation in the data set. There are no absolute differences between the respective groups, and all trends seem to be largely individual. This is reflected in the overlapping HPDI on the group level for each position. Combining the interpretations of the Figures 4–7, we come to the conclusion that no absolute group-specific patterns emerge. Rather, the variation seems to be driven by other, unknown factors. There are some small exceptions to this conclusion, namely that in utterance-final position, monolingual Turkish speakers seem to produce more DM's while some speakers from the US heritage group produce none or close to none, and in utterance-final position, where there are some notable individuals from both heritage groups that produce many DMs, while the remaining speakers from all three groups vary interchangeably.

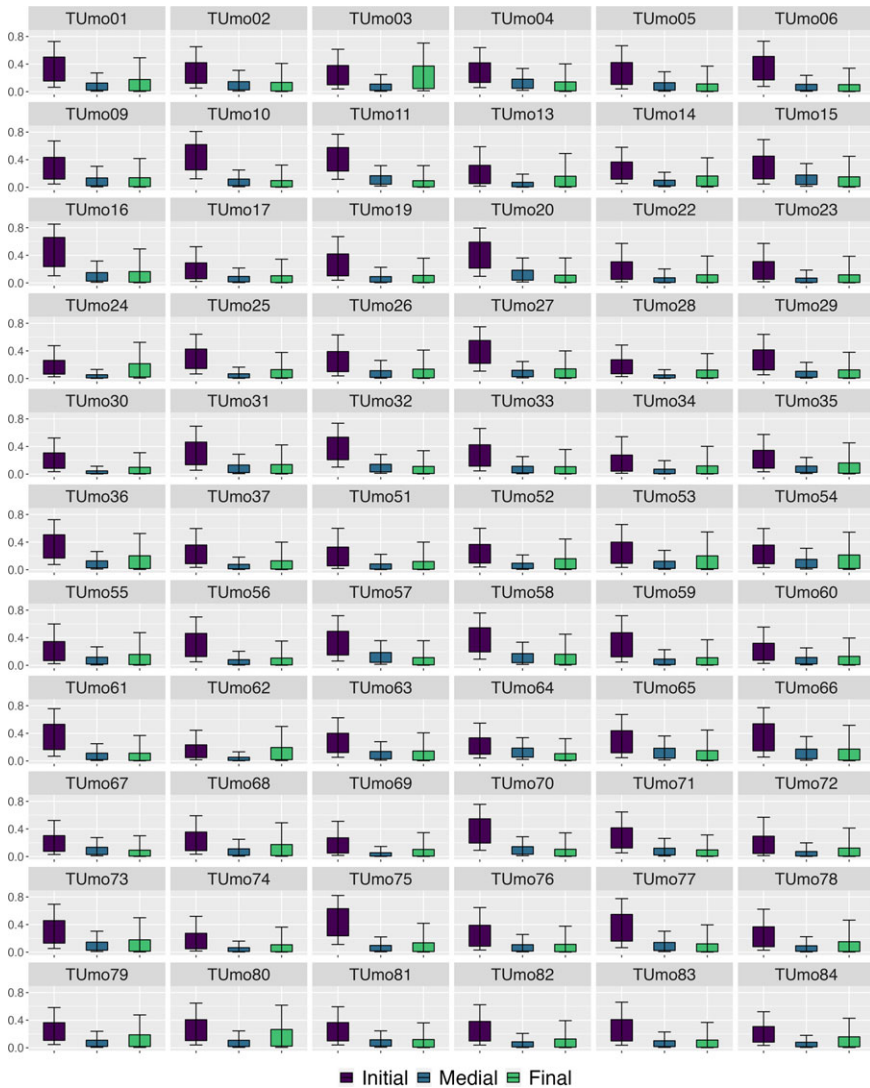


Figure 5. Speaker-level effects for the “Turkey” group.

Discussion

This study investigated two main issues in discussions around heritage language bilingualism. First, we asked whether there are emergent differences in the frequency of production of a linguistic phenomenon (DMs) when we compare heritage and monolingual groups. This is a standard procedure in heritage language research which we critically reevaluated aiming to question a monolingual baseline norm. Assuming that DMs are part of every speakers’ repertoire, we analyzed the use of DMs with regard to various factors such as country of elicitation (group), utterance position and length, and register. This analysis was not limited to

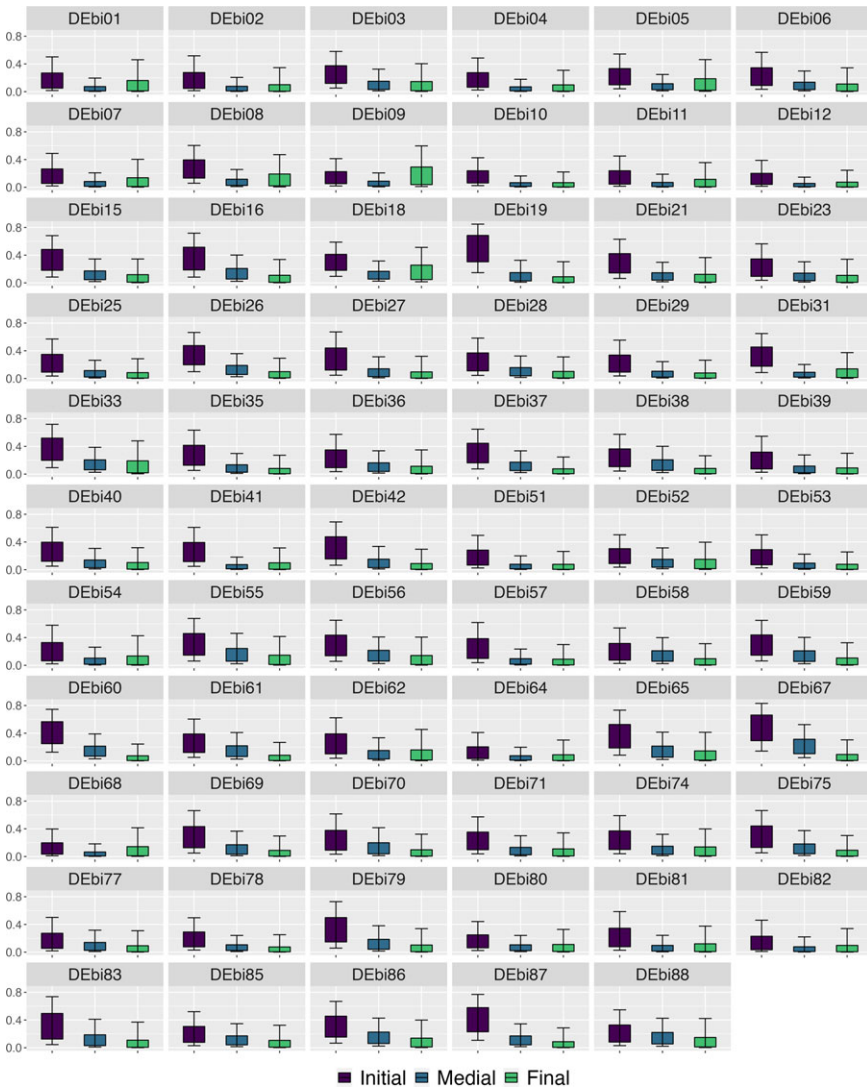


Figure 6. Speaker-level effects for the “Germany” group.

the three forms that we presented as prominent examples above (yani, şey, and işte) and instead includes more than a hundred DMs in total. Our results confirmed the hypothesis that all of these factors affect the production of DMs and FMs. In the genre of heritage linguistics, the focus is then more often on the country of elicitation variable which is generally framed as a grouping variable. The contrast between heritage and monolingual groups is taken as an indicator for novel linguistic developments. The heritage speaker groups in Germany and the USA indeed show marginal diverging patterns of DM use in comparison to the DM use across all three varieties of Turkish here. However, the differences between the



Figure 7. Speaker-level effects for the “US” group.

groups seem to be quite small, and no pattern is consistent across all speakers of a group. This seems to be confirmed by the very small differences in predictive performance of both models, measured by the PSIS-loo information criterion. Instead, the heterogeneity, that is, the individual variation among each group stands out as a much more important finding.

The general observations for utterance position are in line with the literature on DMs, which assumes that they are most likely to occur in utterance-initial positions (Fraser, 1990, 2006). In particular, the effect of utterance position taken together with the group variable reveals an interesting tendency that we might expect for

bilingual heritage speakers: In utterance-medial position, monolingual speakers are overall slightly less likely to produce discourse makers in comparison to the bilingual groups. From a speech planning perspective, utterance-medial DMs indicate micro-planning events like lexical retrieval. Given that bilinguals are generally found to be slower in lexical retrieval than monolinguals (Finkbeiner et al., 2006; Sullivan et al., 2018), it would be plausible that they tend to produce slightly more utterance-medial DMs to bridge the additional time in lexical retrieval. Here, monolinguals do not form a baseline of a language, but the fact that the brain can access lexical items from more than one language's lexicon is expressed from a neurocognitive perspective. In utterance-final position, no clear difference emerges on the group level. It stands out that there do not seem to be any monolingual speakers among the least likely to produce discourse makers, but the HPDI largely overlap due to the low amount of overall observations of DMs in this position.

The second main focus of this study is the speaker effect which measures the individual variation within the groups. Our results show a much bigger variation in the speaker than in the group-level factor. This is in line with our second hypothesis, where we predicted that speaker variation will be a more influential factor compared to the grouping by countries. By shifting our focus to this aspect of the study, we are able to question the framing of monolingual speakers as a homogeneous group which can serve as a baseline. We were able to show that there is large variation in all the groups. Other parameters of our model such as register, utterance length, and position are more important for the use of DMs.

If we add to this the predictive comparison of the grouped and the ungrouped model which showed no predictive benefit of adding the group component to our model, we can conclude that individual variation has a more important role than the variation between groups. This makes it impossible to speak of a general divergence in the heritage speaker group. Therefore, it cannot be assumed that monolinguals are an adequate and homogeneous ground for comparison. Variation must always be treated as a main component of any heritage language study.

In this regard, we argue for more studies that are conceptualized in a way that allow to regard bilingual heritage speakers as native-like too and include them as part of a native speaker continuum, as has recently been suggested by Wiese et al. (2022) and Rothman and Treffers-Daller (2014). It is important to stress, however, that nativeness is always a construct, and many speakers who have years of experience in a language can also be part of that continuum. Otherwise, language becomes something that is solely part of inheritance. As psycholinguists, we should be aware of these circumstances and carefully avoid the reproduction of stereotypes in our studies.

We make room to overthink if and how much we give weight to the group variable in applied bi- and multilingualism research. We echo Serratrice (2020)'s query of why it is useful or whether it is even necessary to compare heritage language acquisition to that of monolingual baselines. Following her analogy, we stress that it also would not be adequate to compare a Turkish speaking group in Istanbul with another group in rural Sivas (Central East Turkey) just because the former group speaks the accepted standard variety of the language. Heritage speakers acquire languages in different contexts and framing their language outcomes as “divergent” or “incomplete” (Polinsky & Scontras, 2020b) can be

problematic. If we as researchers put monolinguals as a baseline for heritage language acquisition, we may (unintentionally) support cultural narratives that label a homeland variety of a language as the norm. As research has shown that these narratives might effect levels and degrees of heritage language anxiety and insecurity (Sevinç & Dewaele, 2018), careful considerations should be made in the designing process of applied psycholinguistic research. We acknowledge that this is an ongoing discussion in the field (see, e.g., Domínguez et al., 2019 and commentaries) where a parting line between theoretically valuable terminological descriptions and wider societal implications of the terminology are in conflict with each other. Previously Bayram et al. (2019) as well as in this special issue Rothman et al. (2022), among others, pointed out that even the arguments for incompleteness do not hold if the individual-level grammars of speakers are considered. While on a group-level heritage speakers might perform by producing fewer target-like grammatical structures that could be interpreted as indicating “incomplete acquisition,” on an individual level the target-like structure is often produced at least once by every speaker which indicates that they have indeed acquired the grammatical structure under question. Therefore, it is not just a terminological but moreover a methodological question that is at stake.

A good alternative for researchers that Serratrice (2020) puts forward are studies that take heritage speakers’ caregivers language as a baseline to measure heritage language attainment and development as studies conducted by Paradis and Navarro (2003) and more recently Coşkun Kunduz and Montrul (2022) illustrate. Polinsky and Scontras (2020a) support this idea and conclude that heritage language research needs to consider heritage speakers’ input in future research to understand heritage language acquisition better. Another strategy would be to view investigations into heritage and monolingual speakers of a language as studies of certain varieties of the language following ideas in variationist sociolinguistics. Such a framing would also diverge from a monolingual baseline and at the same time allow interesting comparisons of equally valid and native varieties of a language.

Whether monolingual comparisons are necessary or not highly depends on the research question at hand. When they are included, they can probably not serve as a “baseline,” as that presents a *comparative fallacy* (Rothman et al., 2022). In this study, we did include a comparison between three groups of native Turkish speakers, two of which were framed as bilingual heritage groups and one as a monolingual homeland group. We achieved this by regressing group as a random effect and not as a main predictor variable. Our aim was to investigate if and how speaker variation overwrites group variation. And indeed we find that speaker variation should be the most crucial factor when investigating DM and FM use in three groups of Turkish native speakers. Most importantly, our results show a strong speaker variation which indicates that discourse and FMs vary mostly based on individual speakers’ patterns. Beyond this, we have learned that overall, utterance length and register are the most important factors that influence DM use. In a one sentence summary, we can say that highly informal settings and long utterances will facilitate the use of DMs.

Given that we as researchers have numerous easily accessible and open-source strategies to capture and analyze individual variation, it seems outdated to simply compare group means. An equally questionable idea would be to just add varying effects to a model without further consideration of these in the analysis. They must

also be interpreted carefully in relation to the main effects that guide the study. Given the scope of this study, we tried to do this with special focus to the group-level effect. This pushed us to shift the perspective from a heritage versus monolingual comparison, to a picture that allows variation between speakers, and interprets all other effects accordingly.

Conclusion

In this study, we explored the frequency in which DMs occur in different utterance positions. A major aspect of our study was to see if the patterns in the data are driven more by individual differences within groups than between groups. Regarding the use of specific DMs, it seems important that further research investigates the many functions even a single DM can have and how these functions are shaped in heritage varieties of languages.

By focusing on variation across groups rather than between groups, our study design explores prevalent monolingualism in our field's research practices and questions common heritage language research procedures. While it remains important to investigate the relation between the sociological construct of heritage speakers with the psycholinguistic reality in language production and comprehension, this approach allows us to unravel the unseen individual linguistic variation in heritage, mono-, and multilingual speakers. In this processes, we should be careful not to overlook similarities or differences that might exist between groups⁶, but at the same time we should aim to carefully take into account other factors that might drive variation. In an attempt to explore the role of individual variation using a new empirical sense, we found that the grouped and ungrouped model show only small differences in predictive performance. Further, no strong pattern can be generalized for any speaker group. This indicates that between-group comparison are not particularly meaningful when investigating DM use on the scale that we did here. Other factors such as register, utterance length, and utterance position within the sentence were much more meaningful in this regard.

In light of our findings and recent discussions about novel approaches in heritage bilingualism research (Cabo & Rothman, 2012; DeLuca et al., 2019; Luk, 2022; Rothman et al., 2022), we emphasize that future studies should carefully consider if any sort of baseline or group comparison is needed in their study. In particular, we recommend to explore if the group factors, which generally assumes some level of homogeneity in the group, is empirically there and meaningful when the within group variation is taken into account.

Replication package. All research materials, data, and analysis code are available on Github (<https://github.com/Tarotis/exploring-individual-variation-in-turkish-heritage-speakers-complex-linguistic-productions>) and published on Zenodo (<https://www.doi.org/10.5281/zenodo.7838068>).

Competing interest. The authors declare none.

Notes

1 We are grateful to an anonymous reviewer who pointed out that a clearer terminological setting here improves the argument that we present.

- 2 We are grateful to an anonymous reviewer for pointing out this additional argument which is crucial in the context and for the argumentation in this paper.
- 3 <https://psyarxiv.com/t4mdj/>
- 4 Further information about the corpus can be found at <https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/rueg/rueg-corpushttps>
- 5 <https://korpling.german.hu-berlin.de/annis3/rueg/>
- 6 We are grateful to a reviewer who shared this conclusion with us in a comment.

References

- Altparmak, A. (2022). An Analysis of Turkish Interactional Discourse Markers ‘şey’, ‘yani’, and ‘işte’. *Journal of Psycholinguistic Research*, 1–34.
- Azar, Z., Özyürek, A., & Backus, A. (2020). Turkish-Dutch bilinguals maintain language-specific reference tracking strategies in elicited narratives. *International Journal of Bilingualism*, 24(2), 376–409. <https://doi.org/10.1177/1367006919838375>
- Baayen, H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bayram, F., Kupisch, T., y Cabo, D. P., & Rothman, J. (2019). Terminology matters on theoretical grounds too!: Coherent grammars cannot be incomplete. *Studies in Second Language Acquisition*, 41(2), 257–264.
- Belz, M., & Odebrecht, C. (2022). Abschnittsweise analyse sprachlicher Flüssigkeit in der Lernaltersprache: Das Ganze ist weniger informativ als seine Teile. *Zeitschrift für germanistische Linguistik*, 50(1), 131–158.
- Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, 125, 104334. <https://doi.org/10.1016/j.jml.2022.104334>
- Brizuela, M., Andersen, E., & Stallings, L. (1999). Discourse markers as indicators of register. *Hispania*, 82 (1), 128–141. Retrieved May 24, 2022, from <http://www.jstor.org/stable/346098>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1). <https://doi.org/10.18637/jss.v080.i01>
- Cabo, D. P. Y., & Rothman, J. (2012). The (il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Applied linguistics*, 33(4), 450–455.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Castro, S., Wodniecka, Z., & Timmer, K. (2022). Am I truly monolingual? Exploring foreign language experiences in monolinguals. *PLoS One*, 17 (3), e0265563.
- Coşkun Kunduz, A. C., & Montrul, S. (2022). Sources of variability in the acquisition of Differential Object Marking by Turkish heritage language children in the United States. *Bilingualism: Language and Cognition*, 1–14.
- Degand, L., Gilquin, G., Meurant, L., & Simon, A. C. (2019). *Fluency and disfluency across languages and language varieties* (Vol. 4). Presses Universitaires de Louvain.
- Degand, L., & Van Bergen, G. (2018). Discourse markers as turn-transition devices: Evidence from speech and instant messaging. *Discourse Processes*, 55(1), 47–71.
- DeLuca, V., Rothman, J., Bialystok, E., & Pliatsikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proceedings of the National Academy of Sciences*, 116(15), 7565–7574.
- Diao, W., & Chen, C. (2021). L2 use of pragmatic markers in peer talk: Mandarin utterance-final particles. *International Review of Applied Linguistics in Language Teaching*, 000010151520200148. <https://doi.org/doi:10.1515/iral-2020-0148>
- Domínguez, L., Hicks, G., & Slabakova, R. (2019). Terminology choice in generative acquisition research: The case of “incomplete acquisition” in heritage language grammars. *Studies in Second Language Acquisition*, 41(2), 241–255.
- Fillmore, C. J. (1979). 5 - on fluency. In C. J. Fillmore, D. Kempler, & W. S.-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-255950-1.50012-3>

- Finkbeiner, M., Gollan, T. H., & Caramazza, A. (2006). Lexical access in bilingual speakers: What's the (hard) problem? *Bilingualism: Language and Cognition*, 9(2), 153–166. <https://doi.org/10.1017/S1366728906002501>
- Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics*, 14(3), 383–398.
- Fraser, B. (2006). Towards a theory of discourse markers. In *Approaches to discourse particles* (pp. 189–204). Brill.
- Furman, R., & Özyürek, A. (2007). Development of interactional discourse markers: Insights from Turkish children's and adults' oral narratives. *Journal of Pragmatics*, 39(10), 1742–1757.
- Gabry, J., Simpson, D. P., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182. <https://doi.org/10.1111/rssa.12378>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC. <https://doi.org/10.1201/b16018>
- Goschler, J., Schroeder, C., & Woerfel, T. (2020). *Convergence in the encoding of motion events in heritage Turkish in Germany*. Studies in Turkish as a Heritage Language, ed. F. Bayram. John Benjamins Publishing Company (pp. 87–103).
- Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. <https://doi.org/10.3366/cor.2015.0068>
- Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, 36 (1), 3–15.
- House, J. (2013). Developing pragmatic competence in English as a lingua franca: Using discourse markers to express (inter) subjectivity and connectivity. *Journal of Pragmatics*, 59, 57–67.
- Iefremenko, K., Schroeder, C., & Kornfilt, J. (2021). Converbs in heritage Turkish: A contrastive approach. *Nordic Journal of Linguistics*, 44(2), 130–154. <https://doi.org/10.1017/S0332586521000160>
- Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.
- Kruschke, J. K. (2015). *Doing Bayesian data analysis*. Elsevier LTD. Retrieved from https://www.ebook.de/de/product/22836901/john_k_kruschke_doing_bayesian_data_analysis.html
- Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5(10), 1282–1291. <https://doi.org/10.1038/s41562-021-01177-7>
- Kupisch, T., Belikova, A., Özçelik, Ö., Stangen, I., & White, L. (2017). Restrictions on definiteness in the grammars of German-Turkish heritage speakers. *Linguistic Approaches to Bilingualism*, 7(1), 1–32. <https://doi.org/10.1075/lab.13031.kup>
- Kurz, A. S. (2021, April 25). Multilevel models and the index-variable approach. Retrieved May 31, 2022, from <https://solomonkurz.netlify.app/post/2020-12-09-multilevel-models-and-the-index-variable-approach/>
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49 (4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Luk, G. (2022). Justice and equity for whom? Reframing research on the “bilingual (dis)advantage”. *Applied Psycholinguistics*, 1–15. <https://doi.org/10.1017/S0142716422000339>
- Makowski, D., Ben-Shachar, M., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>
- May, S. (2020). Rethinking the principle of linguistic homogeneity in the age of superdiversity. In *Language, Nations, and Multilingualism* (pp. 37–53). Routledge.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.
- Montrul, S. (2002). Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and Cognition*, 5(1), 39–68.
- Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge University Press.
- Montrul, S. (2018). Heritage language development: Connecting the dots. *International Journal of Bilingualism*, 22(5), 530–546.
- Nagy, N., & Gadanidis, T. (2021). Heritage language variation and change – how complex is it? *Heritage Language Journal*, 18(2), 1–27. <https://doi.org/https://doi.org/10.1163/15507076-12340012>
- Nicenboim, B., & Vasisht, S. (2016). Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistics Compass*, 10(11), 591–613. <https://doi.org/10.1111/lnc3.12207>

- Niebuhr, O., & Fischer, K. (2019). Do not hesitate!—unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. *Interspeech*, 544–548.
- Oikonomou, D., Rizou, V., Bondarenko, D., Özsoy, O., & Alexiadou, A. (2022). Scalar and counterfactual approximatives: Investigating heritage Greek in the USA and Germany. *Languages*, 7(1). <https://doi.org/10.3390/languages7010011>
- Otcu, B. (2010). Heritage language maintenance and cultural identity formation. *Heritage Language Journal*, 7(2), 273–298.
- Özsoy, O., Iefremenko, K., & Schroeder, C. (2022). Shifting and expanding clause combining strategies in heritage Turkish varieties. *Languages*, 7(3). <https://doi.org/10.3390/languages7030242>
- Paradis, J., & Navarro, S. (2003). Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: What is the role of the input? *Journal of Child Language*, 30(2), 371–393.
- Polinsky, M. (2006). Incomplete acquisition: American Russian. *Journal of Slavic Linguistics*, 191–262.
- Polinsky, M., & Kagan, O. (2007). Heritage languages: In the ‘wild’ and in the classroom. *Language and Linguistics Compass*, 1(5), 368–395.
- Polinsky, M., & Scontras, G. (2020a). A roadmap for heritage language research. *Bilingualism: Language and Cognition*, 23(1), 50–55.
- Polinsky, M., & Scontras, G. (2020b). Understanding heritage languages. *Bilingualism: Language and Cognition*, 23(1), 4–20. <https://doi.org/10.1017/S1366728919000245>
- Poplack, S., & Sankoff, D. (1984). *Borrowing: The synchrony of integration*.
- Putnam, M. T., & Sánchez, L. (2013). What’s so incomplete about incomplete acquisition?: A prolegomenon to modeling heritage language grammars. *Linguistic Approaches to Bilingualism*, 3(4), 478–508.
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reed, M. (2000). He who hesitates: Hesitation phenomena as quality control in speech production, obstacles in non-native speech perception. *Journal of Education*, 182(3), 72–97.
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., Hao, J., Kolb, N., Kubota, M., Kupisch, T., et al. (2022). Monolingual comparative normativity in bilingualism research is out of “control”: Arguments and alternatives. *Applied Psycholinguistics*, 1–14. <https://doi.org/10.1017/S0142716422000315>
- Rothman, J., & Treffers-Daller, J. (2014). A prolegomenon to the construct of the native speaker: Heritage speaker bilinguals are natives too! *Applied Linguistics*, 35(1), 93–98. <https://doi.org/10.1093/applin/amt049>
- Sankoff, G., Thibault, P., Nagy, N., Blondeau, H., Fonollosa, M.-O., & Gagnon, L. (1997). Variation in the use of discourse markers in a language contact situation. *Language Variation and Change*, 9(2), 191–217. <https://doi.org/10.1017/S0954394500001873>
- Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*. <https://doi.org/10.1037/met0000472>
- Schmid, M. S., & Fägersten, K. B. (2010). Disfluency markers in L1 attrition. *Language Learning*, 60(4), 753–791.
- Schroeder, C., & Küppers, A. (2016). Türkischunterricht im deutschen Schulsystem: Bestandsaufnahme und Perspektiven. In *Bildung in transnationalen räumen* (pp. 191–212). Springer.
- Schührer, S. (2018). Türkestämmige Personen in Deutschland: Erkenntnisse aus der Repräsentativuntersuchung “Ausgewählte Migrantengruppen in Deutschland 2015“ (RAM).
- Serratrice, L. (2020). What counts as the baseline in child heritage language acquisition? *Bilingualism: Language and Cognition*, 23(1), 46–47. <https://doi.org/10.1017/S1366728919000518>
- Sevinç, Y., & Dewaele, J.-M. (2018). Heritage language anxiety and majority language anxiety among Turkish immigrants in the Netherlands. *International Journal of Bilingualism*, 22(2), 159–179.
- Shadrova, A., Linscheid, P., Lukassek, J., Lüdeling, A., & Schneider, S. (2021). A challenge for contrastive L1/L2 corpus studies: large inter- and intra-individual variation across morphological, but not global syntactic categories in task-based corpus data of a homogeneous L1 German group. *Frontiers in Psychology*, 12, 716485. <https://doi.org/10.3389/fpsyg.2021.716485>

- Shin, N. (2022). Structured variation in child heritage speakers' grammars. *Language and Linguistics Compass*, 16(12), e12480. <https://doi.org/10.1111/lnc3.12480>
- Shriberg, E. (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169. <https://doi.org/10.1017/S0025100301001128>
- Simmons-Mackie, N. N., & Damico, J. S. (1996). The contribution of discourse markers to communicative competence in aphasia. *American Journal of Speech-Language Pathology*, 5(1), 37–43.
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Sullivan, M. D., Poarch, G. J., & Bialystok, E. (2018). Why is lexical retrieval slower for bilinguals? evidence from picture naming. *Bilingualism: Language and Cognition*, 21(3), 479–488. <https://doi.org/10.1017/S1366728917000694>
- Tagliamonte, S. (2005). So who? like how? just what?: Discourse markers in the conversations of young Canadians [Approaches to Spoken Interaction]. *Journal of Pragmatics*, 37(11), 1896–1915. <https://doi.org/https://doi.org/10.1016/j.pragma.2005.02.017>
- Tocaimaza-Hatch, C. C. (2018). A comparison of formal register through lexical choices in heritage and second language speakers of Spanish. *Linguistics Journal*, 12(1).
- Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, 59(5), 1311–1342. <https://doi.org/10.1515/ling-2019-0051/html>
- Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103, 151–175. <https://doi.org/https://doi.org/10.1016/j.jml.2018.07.004>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved \hat{R} for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2). <https://doi.org/10.1214/20-ba1221>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wiese, H. (2020). Language Situations: A method for capturing variation within speakers' repertoires. *Methods in Dialectology*, 16, 105–117.
- Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Martynova, M., Pashkova, T., Rizou, V., Schroeder, C., Shadrova, A., Szucsich, L., Tracy, R., Tsehaye, W., Zerbian, S., & Zuban, Y. (2022). Heritage speakers as part of the native language continuum. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.717973>
- Winter, B., & Grice, M. (2021). Independence and generalizability in linguistics. *Linguistics*, 59(5), 1251–1277. <https://doi.org/10.1515/ling-2019-0049>
- Yagmur, K. (2011). Does Ethnolinguistic Vitality Theory account for the actual vitality of ethnic groups? A critical evaluation. *Journal of Multilingual and Multicultural Development*, 32(2), 111–120.
- Yılmaz, E. (2004). *A pragmatic analysis of Turkish discourse practices: Yani, işte and şey*.

Cite this article: Özsoy, O. and Blum, F. (2023). Exploring individual variation in Turkish heritage speakers' complex linguistic productions: Evidence from discourse markers. *Applied Psycholinguistics*. <https://doi.org/10.1017/S0142716423000267>