Chapter 2

Tapping into speakers' repertoires: Elicitation of register-differentiated productions across groups

Heike Wiese[a], Annika Labrenz[a] & Albrun Roy[a]

[a]Humboldt-Universität zu Berlin

This chapter addresses the challenge of tapping into speakers' repertoires to elicit ecologically valid data. It describes a set-up that draws on the LangSit method (Wiese 2020) and has been implemented by the RUEG group to elicit data as a basis for the RUEG corpus (cf. Shadrova et al. 2025 [this volume]). Elicitations yielded comparable productions from informal and formal, spoken and written situations across bilingual and monolingual speakers, adults and adolescents, and heritage and majority languages. We discuss strengths and limitations of the set-up and present analyses that illustrate two key strengths of this method: (1) it is successful in eliciting register-differentiated data across different communicative situations, and (2) this data is naturalistic in the sense that it follows the patterns of spontaneous, non-elicited language productions.

1 Introduction

In this chapter, we discuss a method of elicitation that addresses the challenge of tapping into speakers' repertoires to elicit ecologically valid data: we present the set-up that has been implemented by the RUEG group[1] to elicit the data for

[1]Research Unit "Emerging Grammars in Language Contact Situations: A Comparative Approach"; see Wiese et al. (2025 [this volume]) and https://hu.berlin/rueg (last accessed October 23rd, 2023). The corpus created by the group, the open-access RUEG corpus, is described in Shadrova et al. (2025 [this volume]) and can be accessed via https://hu.berlin/rueg-corpus (last accessed October 23rd, 2023).

Heike Wiese, Annika Labrenz & Albrun Roy. 2025. Tapping into speakers' repertoires: Elicitation of register-differentiated productions across groups. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 33–67. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775159

the RUEG corpus. We show that this kind of set-up yields production data that is elicited and thus comparable across different settings and groups, but at the same time representative of speakers' natural behaviour in different communicative situations.

When investigating natural languages, it is important to get a realistic picture of language users' practices and competences. Language use always involves different situation-dependent choices from a broader linguistic repertoire, and accordingly we should take into account such variation. To get a realistic picture of linguistic practices and language competence, we cannot confine ourselves to, say, formal situations associated with standard language. Rather, we need to tap into broader speaker repertoires and cover different registers. At a minimum, this should include language use in informal as well as formal communicative situations.[2]

While this might seem obvious, it is something that often gets overlooked in research. One area where we see this as a desideratum is research on heritage-language speakers. These are bi- or multilingual speakers who grew up in a bi- or multilingual home with a minority language in addition to the majority language of the larger society (see Introduction).

Heritage languages are often used primarily in informal domains: they are acquired in the family and often remain associated with informal communicative situations. In contrast, languages spoken as a societal majority language will also be used in more formal situations, and the acquisition of formal registers of such languages will be further supported as part of schooling (see also Kupisch & Rothman 2018).

Yet, heritage language research often looks at data elicited under lab conditions that are favourable to more formal language, and compares heritage speakers' data to that of monolinguals who speak the language as a majority language. In such lab settings, monolinguals will choose formal registers, while heritage speakers might use more informal versions if formal registers of the heritage language are not part of their repertoire. This, then, brings in an additional factor: informal patterns in bilinguals are measured against formal ones in monolinguals. Hence, if we do not want to compare apples and oranges, we need to take into account informal as well as formal communicative situations across groups, in monolinguals and multilinguals alike.

[2]Following the terminology introduced in the Introductory chapter (Wiese et al. 2025 [this volume]), we understand communicative situations as the settings of a communication, and registers as the linguistic counterparts of communicative situations, that is, the language use associated with different communicative situations.

Furthermore, it can be valuable to include both spoken and written language. Heritage speakers typically do not acquire literacy in their heritage language as part of formal schooling, so it might depend on their media consumption (books, newspapers) in how far they acquire formal-written registers. Informal-written registers of the heritage language might be used in digital social media, for instance transnationally in interactions within extended families, but this might be not as common as, say, text messaging in the majority language. Including spoken vs. written as well as formal vs. informal communicative situations can therefore allow us to tap into and tease apart different relevant domains of speakers' linguistic resources.

Taking into account such register distinctions provides us with ecologically more valid data on both monolingual and bi-/multilingual speakers' competences.[3] Furthermore, comparing both groups on equal grounds gives us a better picture of heritage language characteristics, since it allows for a better assessment: What phenomena are really specific to heritage languages, language contact or multilingualism? What phenomena, in contrast, might actually be found across speaker groups if we do not restrict ourselves to, say, formal situations?

This is not only relevant for the heritage language, but also for another part of heritage speakers' resources, namely the majority language. Heritage speakers who grow up in a country with a different majority language are often not acknowledged as part of the core speaker group of that language, and their linguistic practices are then approached from a deficit perspective. Such effects of a widespread monolingual habitus in the Global North are evident not only in public discourse, but also in linguistic papers.[4] As a result, patterns of linguistic variation in informal or spoken registers might not be recognised as such, but attributed to bi-/multilingualism. If we take into account register-differentiated use in bilinguals and monolinguals alike, such misattributions can be avoided. This can provide new insights into native grammars, their dynamics in multilinguals, and their variability across registers (see also Wiese 2013, Kupisch & Polinsky 2022 on the special dynamics of multilingual settings).

Wiese et al. (2022) discuss cross-linguistic examples for the kind of findings that such an approach can yield. Tapping into speakers' broader repertoires re-

[3] In corpus-linguistic research, the problem of the authenticity (or naturalness) of corpus data was discussed as early as the 1990s: one of the questions was what kind of texts (e.g. informal/-formal; oral/written; literary/ordinary) should be included in a sample so that a corpus reflects "authentic" language (e.g. Sinclair 1991) or is representative of a particular language use (e.g. Biber 1993).

[4] See Wiese et al. (2022) for a critique; Adli & Guy (2022) for a call to broaden our perspective and reduce the bias towards Western societies.

vealed some unexpected patterns not just in heritage speakers, but also in monolinguals. We found, for instance, noncanonical word order patterns, such as new referents in post-verbal position in Turkish[5] or the placement of two constituents before the finite verb in main declaratives in German.[6] Both of these patterns are ruled out by established grammatical descriptions of Turkish and German, respectively. Yet, we found them to be systematically available in native grammars. In majority language settings, such patterns were associated with informal and/or spoken language, indicating their register-related status. Hence, when heritage speakers use such noncanonical patterns in more formal settings, this points to register levelling rather than novel patterns.

Such findings underline the need to cast our net wider and target ecologically valid data in a broader range of communicative situations: had we focused on formal situations alone, such patterns might only have shown up in heritage speakers and might then have been misattributed to bilingualism. In contrast, if we tap into actual language use and look at broader repertoire data from multilinguals and monolinguals alike, then this can reveal interesting similarities of noncanonical patterns and inform us on the dynamics of native grammars (see also Bayram et al. 2019 on the problem of using an idealised monolingual standard as a yardstick).

However, it is notoriously difficult to gain ecologically valid data that is representative of speakers' natural behaviour and covers not only formal, but also informal settings. The empirical basis for much of heritage language research so far has often been limited to data elicited under lab conditions (see also Montrul 2015: Chapter 6, Polinsky 2018: Chapter 3 for examples). Common set-ups are, for instance, ones that use acceptability, sentence completion, or picture matching tasks. This kind of elicitation has two major advantages. For one, it yields controlled data that is suitable for targeting specific linguistic phenomena that have been of interest for the analysis, for instance, certain morphological or syntactic patterns in a heritage language. Second, it yields comparable data across speaker groups, and it has typically been employed to compare heritage speakers to monolinguals (but see Rothman et al. 2023 for a call to further reflect on the choice of comparison groups in heritage language research). However, such elicited data also has a major disadvantage: it is not representative of speakers' natural behaviour across different communicative situations (see also Aalberse et al. 2019: Chapter 5 for a discussion of this problem). Such data reflects only those responses in a somewhat artificial setting that will favour formal language

[5]Wiese et al. (2022); for detailed discussion of this pattern see also Schroeder et al. (2024).

[6]Wiese et al. (2022); see also Wiese & Müller (2018) for an analysis of verb-third in German.

use, and as we have argued above, this can mean missing out on important patterns in multilinguals and monolinguals alike.

A type of method that avoids artificial settings and has also been employed in minority language research is the collection of spontaneous data, for instance through speakers' self-recordings (see Aalberse et al. 2019: Chapter 5 for examples). This kind of data can be regarded as the gold standard from the point of view of ecological validity, but it is much more difficult and time-consuming to collect. Furthermore, since such data is spontaneous and thus by its very nature not controlled, it does not offer the same advantages as elicited data: specific linguistic patterns might just not occur (at all) in the database, and it does not yield comparable productions across speakers and speaker groups. Additionally, in an uncontrolled setting, the data might only represent a specific part of a speaker's repertoire, e.g., informal conversations with friends, but not, say, formal registers used with strangers.

A kind of methodology that addresses these challenges and combines the advantages of controlled and spontaneous data is the elicitation of naturalistic data. This is data that is representative of speakers' natural behaviour, but at the same time is elicited and thus can be more controlled, making it suitable for targeted and comparative investigations. In what follows, we present a setup that achieves this and provides data that covers speakers' behaviour across different communicative situations. This is the "Language Situation" (LangSit) method. In Section 2, we describe this methodology and its implementation by the RUEG group to yield comparable data across a broad range of different settings and speakers, which supported such empirical findings on heritage speakers and monolinguals as mentioned above. Section 3 will present analyses that illustrate two key strengths of this method: (1) it is successful in eliciting *register-differentiated* data across formal and informal, written and spoken communicative situations, and (2) this data is *naturalistic* in the sense that it follows the patterns of spontaneous, non-elicited language productions. In the final section (Section 4) we conclude this chapter and discuss the implications of our results.

2 Naturalistic elicitations across groups based on the LangSit method

2.1 LangSit: General features

The LangSit method has been used to elicit naturalistic data across different communicative situations in a range of different studies, in particular on multilingual

settings, but not restricted to these, and it provides the set-up for RUEG's empirical basis, the RUEG corpus (see Shadrova et al. 2025 [this volume] and Klotz et al. 2024). In this subsection, we describe its core features; for a more detailed discussion (including an overview of previous applications) see Wiese (2020) and the LangSit website (https://hu.berlin/LangSit).

The main idea underlying the LangSit method is to encourage speakers to playact *natural communication* in *different situations*. The first part aims at eliciting naturalistic language, the second part targets register-differentiated productions, allowing us to cover different parts of speakers' repertoires. Finally, in order to make productions *comparable* across the different communicative situations, speakers are asked to talk about the same event in each situation.

This event is presented to participants through a nonverbal stimulus (to prevent linguistic priming), for instance a photo story or a short video. In order to make the setting ecologically valid, the event should be interesting enough to motivate people to talk about it. For instance, for a corpus of Namibian German (the DNam Corpus[7]), we showed participants a photo story where a car bumped into a trolley on a shopping mall car park, causing a woman to fall down and spill the contents of her bag (cf. Wiese et al. 2017, Zimmer et al. 2020).

Participants are asked to imagine being a witness to the event and then to playact telling different people about this. Through the choice of different interlocutors, researchers can specify different communicative situations. For instance, asking speakers to tell a friend or a relative about what they saw will simulate an informal situation; asking them to tell a police officer or a teacher simulates a formal situation. Further domains of speakers' repertoire can be targeted by manipulating additional aspects of the language production, for instance the mode by asking participants to write or to speak, or the monologic vs. dialogic communicative character by asking them to leave a message for someone or to speak to each other.

To elicit this kind of language production, participants are familiarised with the event and then asked to act out telling someone about this. For instance, in the informal situation, they are told something along the lines of "Imagine you just witnessed this event. What would you do? You might, for instance, call your friend and tell her about this, or you might be asked to describe this to a police officer who needs a witness. Let's act this out. Talk to a friend about it. Which friend would you call? Now take your phone and do as if send them a voice message."

[7]See https://www.linguistik.hu-berlin.de/de/institut/professuren/multilinguale-kontexte/korpora/dnam (last accessed 2025-02-13).

The fact that participants are asked to talk to different types of interlocutors that are associated with informal vs. formal communicative situations is advantageous not only because it supports register-differentiated data. It also reduces speakers' reluctance to produce informal language in the setting of an elicitation, since it is clear that this will not reflect badly on their language competence, given that they are asked to produce formal language as well. This focus on different situational choices thus helps speakers to overcome the restraints of standard language ideologies, something that is a notorious obstacle to eliciting informal language use (Wiese 2020). On the other hand, for speakers who might not have formal registers of a heritage language in their repertoire, it can be of help for them to know that they are also asked to produce informal language: the latter allows them to act out in a setting that they associate with their heritage language and where they feel secure to use it. This can reduce the kind of language anxiety that might hold them back and distort their responses in a formal lab setting (see Sevinç & Dewaele 2018 on language anxiety in bilingual speakers).

Finally, the choice of event can guide speakers towards specific topics (e.g., make them talk about a trolley or about someone falling down) and thus control to some degree the kind of linguistic material they use, which can help target specific linguistic domains (e.g. nouns of a certain gender, or motion verbs).

Taken together, the LangSit method is open enough to yield ecologically valid, naturalistic data; it is powerful enough to cover different communicative situations and thus elicit register-differentiated productions; yet it is restrictive enough to control the type of situations and topics and to yield comparable data across speakers, speaker groups, and (contact-linguistic) contexts.

2.2 RUEG's LangSit implementation

The RUEG group employed the LangSit method to collect data for our shared empirical basis. As described in the introductory chapter (Wiese et al. 2025 [this volume]), our overarching goal was to investigate heritage speakers' linguistic competences and the dynamics of language contact. To this aim, we explored the distinctive grammatical and pragmatic characteristics of heritage language use, the role of bilingualism and the impact of the contact situation. To achieve this, we conducted large-scale comparisons across languages, speaker groups, and settings (see Wiese et al. 2025 [this volume]). The close comparability that LangSit affords was hence a major advantage for us. A second advantage was its power to yield naturalistic data, for bilinguals and monolinguals alike.

When implementing the LangSit method, we made a number of specifications to the general set-up, which we describe in the following paragraphs. To ensure uniform data collection, we conducted an on-site training course for all elicitors that also included a video tutorial. All materials we used are fully open-access, available through RUEG's site on the Open Science Foundation (https://osf.io/cm96g/). This includes an overview of the set-up, all stimuli, detailed instructions for elicitors, and the training video for data collection.

We used a *video stimulus* to present the event, in order to give participants a lively and natural impression that would make it easier for them to imagine themselves as a witness. Following previous LangSit implementations, we chose a minor traffic accident, as an event that is notable enough to motivate telling about it, but not so severe as to disturb viewers.

Since we conducted our elicitations in *five countries* (Germany, Greece, Russia, Turkey, and the US), we tried to avoid any elements in this accident that were too specific to work equally well in each country. When piloting the set-up, we had used an accident between a car and a bike at a traffic light. The results of piloting showed that this was too country-specific: among other things, traffic lights look different across countries, and cyclists are less common in some countries than in others. Accordingly, for the main study we set the accident on a car park, with no bike involved. This said, one cannot altogether eliminate the possibility that country-specific differences have an impact, for instance we might find different attitudes and experiences with speaking to the police in different countries or for different speaker groups (for instance, depending on the groups, there might be higher levels of language anxiety in heritage speakers in such communicative situations).

The story is about a woman with a dog on a leash who is loading groceries into her car when a young man and a woman with a baby pram walk towards her from across the car lane. The man plays with a ball that suddenly rolls into the lane towards the woman with the dog. The dog breaks free and runs into the lane towards the ball, making the woman drop her groceries. Meanwhile, two cars have approached; the driver of the first car has to break abruptly for the dog, and as a result the second car bumps into the first car. The two drivers get out and call the police. The young man checks on them and then helps the dog owner retrieve her groceries.

We used the same video across all countries, except for one element: in order to show that the car drivers called the police, there was a close-up of one driver's mobile phone showing the police telephone number, which is different in different countries, so we used country-specific pictures for that.

To cover a broad spectrum within speakers' repertoires, we defined *four communicative situations* that differed along the parameters of formality and mode: informal-spoken, informal-written, formal-spoken, and formal-written. (In-)formality was induced by different *interlocutors*: for the informal productions, participants were asked to play-act telling a friend about the incident; for the formal ones, describing the accident as a witness for the police. Hence, the parameter of formality was operationalised through different interlocutors, with a friend representing a familiar interlocutor in a more symmetric and closer relationship to the speaker, and a police officer representing an unfamiliar interlocutor in a more asymmetric and professional relationship to the speaker. This does not cover all possible aspects that can contribute to the formality vs. informality of a communicative situation, of course, but it captures some key distinctions that are likely to constitute situations of different formality for speakers.

To enhance comparability across spoken and written productions, communication was set up as monologic in both modes. Participants were asked to type texts in the written mode and to leave voice messages in the spoken mode; hence all productions were *non-simultaneous*. In the case of Russian and Greek, we left the choice of script (Latin, Cyrillic, Greek) open for heritage speakers.

In the informal productions, participants sent messages via WhatsApp from a mobile phone provided by us, with auto correction, swiping, and suggestions switched off. For formal-spoken productions, we used the same mobile phone, where we had created a contact "Police Department – eyewitness line"; participants were told to imagine that this was a line for witnesses to leave a voice mail with their reports, and that they had been asked to do this for the accident they had seen. For formal-written productions, participants were asked to type a witness report for the police on a laptop we provided.

Across participants, we *balanced* the *order* of communicative situations, that is, we had a balanced order of informal and formal parts, and of written and spoken elicitations within informal and formal parts.

After elicitations, participants filled in a digital *questionnaire* with questions on their personal details (e.g., age, date and place of birth, educational background, family members), language biographies, language and media use. This information was recorded as a basis for analyses targeting the differential impact such factors might have on language productions.[8]

It is generally challenging to elicit informal productions. As discussed above, the LangSit set-up somewhat mitigates this problem by its emphasis on speakers'

[8]The importance of documenting metadata that might have an impact on language variation (e.g., available metadata on texts and producers) has been emphasised in corpus linguistics for a long time (cf. Sinclair 1991, Granger 2008).

repertoires, which makes it clear for participants that informal language use will not be judged as incorrect or bad. To make it easier for participants to differentiate informal and formal elicitations, there was a break between them. In addition, we included a conversational session (*chit chat*) before the informal elicitation, in order to help participants to get into the spirit of an informal conversation.

We further supported the production of naturalistic informal as well as formal language by creating different settings for informal and formal elicitations, with different surroundings and elicitors. We used two elicitors for each meeting. *"Formal" elicitors* wore a formal outfit (for instance, a suit) and interacted with participants in a distanced and formal manner, including through their way of speaking, for instance, they used the formal *Sie*-form in German (the *vos* variant in the T-V distinction, see e.g. Bilá et al. 2020), addressed participants with their surname, and used language close to the standard. In addition to the elicitation, they also organised all the formalities, e.g., they handed out speaker information forms, consent-forms, and receipts for the reimbursement. *"Informal" elicitors* wore informal clothes (for instance, jeans, a sweatshirt and a basecap), acted in a relaxed and accessible manner, offered cookies and drinks, and used more vernacular language, using participants' first name, and e.g., addressing them with the informal *du*-form (*tu* variant) in German. They also interacted with participants during the chit-chat sessions.

Formal and informal elicitations took place in different rooms designed to further induce (in)formality: the *formal room* was an office room with a desk and no decoration; the *informal room* had comfortable seating and was made to look cosy and informal, for instance, by draping colourful cloths over chairs, putting a candle on the table, using mismatched mugs etc.

Since we were interested in speakers' competences and language use across their repertoire, we included *both languages* for bilingual speakers. This meant that bilingual speakers did two sessions, one in their heritage language and one in the country's majority language. The sessions were at least three days apart. Again, the order was balanced. Each session was conducted in a monolingual mode, with elicitors who used only the respective language (either the heritage or the majority language). This was owed to our research interest in the dynamics within heritage and majority languages, both for bilinguals and monolinguals.

Taken together, this allowed us to collect comparable data targeting speakers' repertoires across speaker groups, countries, and heritage and majority languages. The data was processed and integrated into a shared empirical basis, the *RUEG corpus* (Lüdeling et al. 2024), an annotated, multi-layer and multimodal open-access corpus which is freely available in an open repository (see also Shadrova et al. 2025 [this volume] and Klotz et al. 2024).

As mentioned above, one of the strengths of this database was that it revealed unexpected patterns in monolingual as well as bilingual speakers, allowing for new insights into native grammars. The large-scale comparisons that it made possible enabled us to tease apart different factors underlying linguistic variation and the development of noncanonical patterns, and to pinpoint the role of language contact (see contributions in this volume). In the next section, we evaluate the power of our methodology to yield naturalistic register-differentiated productions: How successful was our method in tapping into speakers' repertoires, that is, did we manage to elicit data that is representative for different registers? And how close is this elicited data to spontaneous language productions under natural conditions, that is, how high is the price we pay for using controlled elicitations that yield comparable data?

In what follows, we arrive at a positive evaluation when addressing these questions: we show that our set-up was successful in eliciting register-differentiated data across formal and informal, written and spoken communicative situations (Section 3.1), and we show that the elicited data is naturalistic in the sense that it follows the patterns of spontaneous, non-elicited language productions (Section 3.2). For the first part, we present comparisons within the RUEG corpus, for the second, we present comparisons of the informal-spoken RUEG corpus data with a corpus of spontaneous WhatsApp voice messages.

For comparisons within the RUEG corpus, we target phenomena at different linguistic levels, namely syntax (Section 3.1.1), discourse organisation (Section 3.1.2), and the lexicon (Section 3.1.3), and show that they vary according to our four communicative situations. Comparisons of our corpus data with spontaneous voice messages indicate similar patterns, and again we show this for a range of phenomena at the levels of syntax (Section 3.2.2), discourse organisation and the lexicon (Section 3.2.3). Analyses of the RUEG data are based on the RUEG-1.0-SNAPSHOT corpus version (2023).

3 Register-differentiated naturalistic productions

3.1 Register distinctions within the RUEG corpus

3.1.1 Syntax

A central domain that has been found to be sensitive to register distinctions is that of syntactic categories, and a prominent example is the distribution of nouns vs. pronouns. As Biber & Conrad (2019: Chapter 3) showed, pronouns tend to be more frequent in situations with a higher degree of interaction between

interlocutors where they refer to referents that the addressee is familiar with or that are present in the communicative situation. In contrast, full noun phrases can convey information that is unknown to the addressee and are therefore more frequent in situations where interlocutors are less familiar with each other.

Accordingly, if our elicitation schema worked out, we should expect a more frequent use of pronouns in the corpus data from spoken and informal elicitations, and a more frequent use of nouns in the corpus data from written and formal elicitations.

To test this, we used the universal part-of-speech tags that the RUEG corpus annotations provide for all five languages (English, German, Greek, Russian, and Turkish).[9] We conducted a corpus query based on the tags "NOUN" and "PROPN" for nouns (common nouns and proper nouns) and "PRON" for pronouns. For comparison of frequencies, we provide normalised frequencies, calculated as occurrences per 100 tokens, where the total number of tokens included all tokens except punctuation marks and filled pauses. We conducted Kruskal-Wallis tests and posthoc pairwise Wilcoxon tests.[10] Figure 1 gives the results for nouns, and Table 1 shows normalised rates of occurrences per 100 tokens for nouns across languages and registers.

The pattern confirms the expected register differences for all languages (">" indicates significant differences of $p < 0.01$):

formal-written > informal-written, formal-spoken > informal-spoken

We hence see an impact of both the parameters we manipulated for register distinctions: formality as well as written mode supported the use of nouns. The two parameters reinforced each other, with the effect that the setting with two positives (formal and written) elicited most nouns, that with two negatives (informal and spoken) the least, and the two conditions that combine a positive and a negative instance (informal and written or formal and spoken) occupy a shared intermediate position. Table 2 gives the results for pronouns.

[9] https://universaldependencies.org/u/pos/ (last accessed: 2025-02-14; see Shadrova et al. 2025 [this volume] for more details)

[10] Note that the standard deviations are not high in this domain, which is advantageous for the comparisons and might not hold for some other linguistic domains. As is frequently mentioned in heritage language research, we often observe a large variability in heritage speaker data (see, e.g., Montrul 2015, Polinsky 2018). Note, though, that this is not restricted to this speaker group: as shown in Wiese et al. (2022), we can find the same or even a higher degree of variability in monolinguals (see also Shadrova et al. 2025 [this volume] on the massive variability in naturalistic language in general).

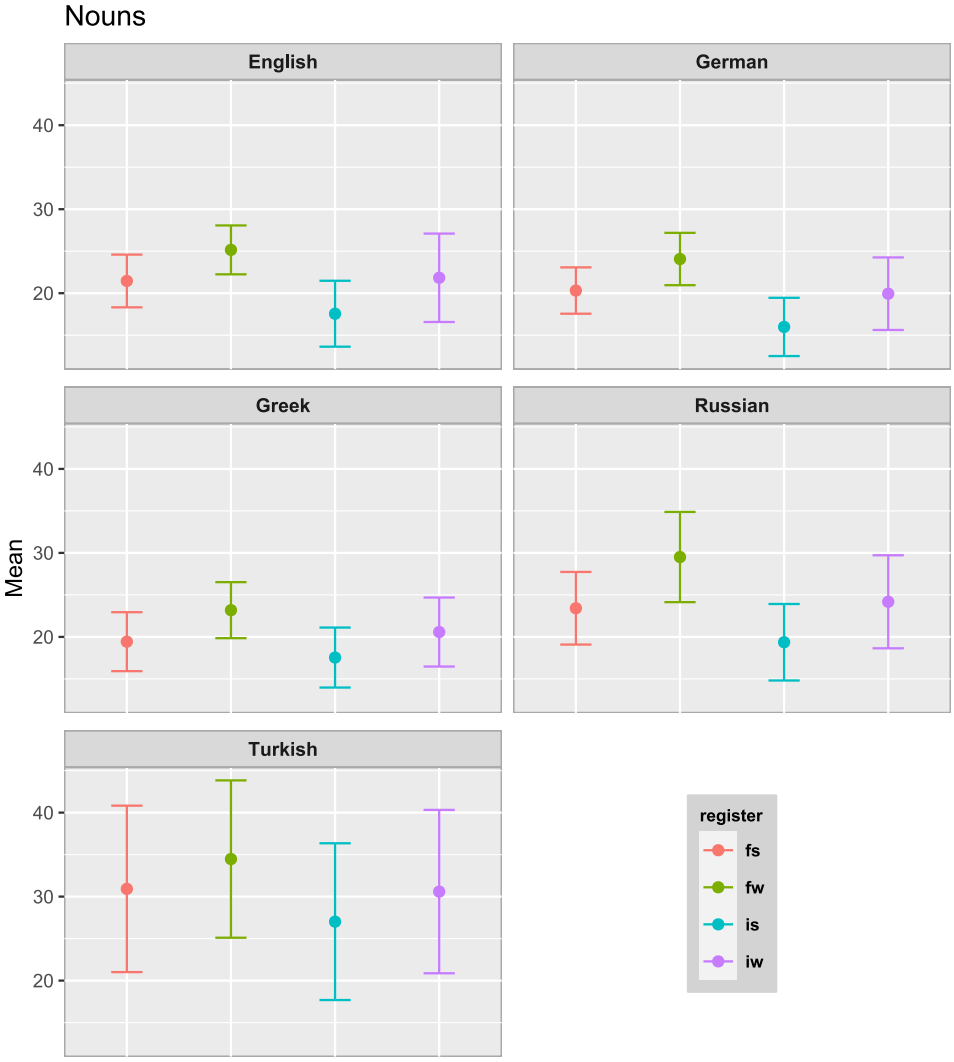

Figure 1: Means (dots) and standard deviations of normalised frequencies of nouns across languages

Table 1: Means of normalised occurrences per 100 tokens ("No.") and standard deviations ("SD") for nouns across languages and registers

	Formal				Informal			
	written		spoken		written		spoken	
Language	No.	SD	No.	SD	No.	SD	No.	SD
English	25.16	2.91	21.46	3.14	21.83	5.27	17.56	3.92
German	24.07	3.12	20.32	2.76	19.94	4.32	16.00	3.47
Greek	23.18	3.33	19.43	3.51	20.58	4.12	17.54	3.57
Russian	29.51	5.37	23.41	4.32	24.18	5.54	19.36	4.55
Turkish	34.47	9.37	30.91	9.91	30.60	9.33	27.02	9.73

Table 2: Means of normalised occurrences per 100 tokens ("No.") and standard deviations ("SD") for pronouns across languages and registers

	Formal				Informal			
	written		spoken		written		spoken	
Language	No.	SD	No.	SD	No.	SD	No.	SD
English	4.49	2.23	7.38	3.05	8.94	3.43	10.05	3.50
German	6.22	2.34	8.02	2.70	8.35	3.35	10.10	3.39
Greek	7.29	3.07	9.23	3.20	8.90	3.62	9.50	3.17
Russian	6.76	3.25	10.05	3.98	9.03	3.92	10.74	3.82
Turkish	4.07	2.66	4.96	3.08	5.66	3.33	6.10	3.85

This pattern confirms our expectation to find least pronouns in the formal-written condition, but unlike the pattern for nouns, we find some cross-linguistic differences for the other conditions. For English, and German, the pattern mirrors that for nouns (">" indicates significant differences of $p < 0.01$):

informal-spoken > formal-spoken, informal-written > formal-written

For Turkish, Greek, and Russian the pattern is similar, but with slight differences: For Turkish and Greek, there is no significant difference between frequencies in the informal-spoken condition and the two intermediate ones (formal-spoken and informal-written):

informal-spoken, formal-spoken, informal-written > formal-written

For Russian, there is no significant difference between frequencies in the formal and informal-spoken modes, but there is one between those and the written conditions as well as between the informal-written and formal-written conditions. This brings mode to the fore as a relevant parameter for the frequent use of pronouns:

informal-spoken, formal-spoken > informal-written > formal-written

The difference between German and English, as opposed to Turkish, Greek, and Russian is probably due to typological differences. Russian, Turkish and Greek are languages characterised by frequent pro-drop (see Özsoy et al. 2025 [this volume][11]). As shown in Table 3, we found notably fewer personal pronouns in Turkish and Greek compared to the other languages. For Russian this holds especially for the written mode. These differences might account for the lack of significant differences between registers regarding the category of pronouns. It seems that in pro-drop languages, pronouns are a less decisive linguistic indicator for register differences.

Table 3: Means of normalised rates of occurrences per 100 tokens of personal pronouns across languages and registers

	Formal		Informal	
	written	spoken	written	spoken
English	5.39	3.74	6.22	7.95
German	4.34	3.21	4.17	4.91
Greek	1.07	0.41	0.24	1.19
Russian	3.42	1.67	2.55	3.77
Turkish	2.36	1.73	1.43	2.03

Across communicative situations, nouns are more frequent overall than pronouns. This is presumably due to the specific communicative task that participants had, namely to retell a car accident where a lot of new referents had to be introduced (several persons, a dog, a ball, cars, a pram etc.). This points to a general challenge of the LangSit method: the topic under discussion can have an

[11]For specific effects of mode and formality on subject drop in Russian and Turkish and on demonstratives in Russian, Turkish, and Greek cf. Özsoy et al. (2025 [this volume]).

impact on the linguistic means that speakers use, in this case the frequency of nouns, and this is something one has to take into account when comparing such corpus data with spontaneous data or data elicited in other set-ups. This is not an issue when we compare LangSit data elicited in the same set-up within our corpus, since we used the same stimulus video throughout, and these comparisons point to systematic patterns of differences between communicative situations.

Taken together, our data hence indicates that our set-up was successful in eliciting register-differentiated data: it constituted different communicative situations that were perceived by participants as suitable for differentiated language use. In this dynamic, (in)formality and mode augmented each other in a way that the formal-written condition was located at one pole of the continuum and the informal-spoken one at the other pole.

As the findings presented in the other chapters show, such register distinctions are also evident at the level of complex syntactic patterns, for instance for clause types in English (Pashkova et al. 2025 [this volume]), left dislocation in German, Russian, and English , verb-second violations in German and OVS vs. SVO word order in Russian (for the last three examples see Bunk et al. (2025 [this volume])).

3.1.2 Discourse organisation

In the domain of discourse organisation, openings are particularly interesting for distinctions between informal and formal registers, since the choice of openings is closely related to the relationship between the speaker/writer and the interlocutor.

Our corpus data is well suited to investigate this: it contains a lot of openings since the task was to retell an incident, and openings are typical for narratives in general. They are usually less frequent in WhatsApp messages between friends (which is what we chose for the informal conditions) because these are often embedded in a longer, continuous communication process. However, this was not the case in our elicitations since speakers had to act out sending a message out of the blue rather than relating to a previous conversation. Accordingly, we find enough openings even in the informal conditions.

Qualitative analyses show clear differences between the data elicited in informal versus formal conditions, and very similar patterns across spoken and written modes, in particular within productions in informal conditions.

(1) through (4) give examples from majority-German, from informal productions in spoken ((1) and (2)) and written mode ((3) and (4)).[12]

[12]The speaker codes are constructed as follows: The first two letters refer to the country of elicita-

(1) Ey [name] du weißt nicht was passiert ist
ey [name] you know not what happened has
'Ey [name] you don't know what happened.' [DEbi71FG_isD]

(2) Hallo meine liebe du weißt nich was schon wieder passiert is
hello my dear you know not what yet again happened has
'Hello my dear, you don't know what happened yet again'
[DEbi16FT_isD]

(3) Jo dicker, du weiss nicht, was ich grad erlebt hab
yo fat.one, you know not what I just experienced have
'Yo dude, you don't know what I just saw' [DEbi04FT_iwD]

(4) [Name]! du weiß nicht was passiert ist
[name]! you know not what happened has
'[Name]! you don't know hat happened' [DEbi85FR_iwD]

(5) through (8) illustrate the contrast to formal productions in majority-German, in spoken (5 and 6) and written mode (7 and 8):

(5) ja guten tag DEbi10MR mein name
yes good day DEbi10MR my name
'Yeah, good afternoon, my name is DEbi10MR' [DEbi10MR_fsD]

(6) Hallo ja aso mein name is DEbi68FR
hello yes so my name is DEbi68FR
'Hello, yes, so, my name is DEbi68FR' [DEbi68FR_fsD]

(7) Sehr geehrte Damen und Herren, nachfolgend finden Sie meinen
very honoured ladies and gentlemen following find you my
Zeugenbericht zum Unfall vom xxx
witness.report on.the accident of xxx.
'Dear Madam/Sir, please find below my witness report on the accident of xxx.' [DEbi04MT_fwD]

tion (DE=Germany, TU=Turkey, GR=Greece, US=USA RU=Russia, TU=Turkey); the next two to the speaker's language background (bi=bilingual, mo=monolingual) followed by speaker number, which also indicates one of two age groups (speakers numbered 1–49 are adults, those with numbers >50 are adolescents); then gender self-identification (F=female, M=male), heritage language (G=Greek, D=German, T=Turkish, R=Russian) or majority language in case of German (D) and English (E) monolinguals; communicative situation (iw = informal-written, is= informal-spoken, fw = formal-written, fs = formal-spoken); language of elicitation (D=German, E=English, G=Greek, R=Russian, T=Turkish).

(8) Sehr geehrte Damen und Herren, meine Fallnummer ist F16
very honoured ladies and gentlemen, my case.number is F16
'Dear Madam/Sir, my case number is F16' [DEmo56FD_fwD]

As the examples illustrate, we find characteristic patterns that differ for data elicited in informal vs. formal conditions.[13] In the informal conditions, the imagined interlocutor is a friend, and participants used an informal greeting, usually followed by a form of address and an attention getter along the lines of "You do not know/believe what just happened" that motivates the narrative. In formal conditions, participants were asked to imagine a police officer as an interlocutor, and in this condition, they used a formal greeting, often gave their name or the case number and then introduced their report.

The marked differences between these patterns indicate register distinctions at the level of discourse organisation: they confirm that the set-up was successful in eliciting distinct discourse openings in settings presented as informal vs. those presented as formal. The findings suggest that participants perceived the communicative situations as different along the parameter of (in)formality and acted accordingly.

3.1.3 Lexicon

Interacting with discourse organisation, we also find differences between openings in informal vs. formal conditions at the lexical level. As the examples above illustrate, participants chose colloquial terms (*jo*, *ey*) in informal openings and addressed the interlocutor with her/his first name, a term of endearment (e.g. *my dear*), terms of address associated with peer-group situations among young people (e.g. *Dicker*), or a nickname. In contrast, in the formal conditions they used formal greetings (*Guten Tag*, *Sehr geehrte Damen und Herren*) that indicate a social distance between interlocutors, and then presented themselves or the case number. An exception is *hallo*, which serves as a neutral term of greeting that is used across communicative situations.

As is discussed in Labrenz et al. (2025 [this volume]), we also find register distinctions within lexical elements at the level of discourse functions (see also Labrenz 2023): the German word *also* shows functional variation across registers that reflects the linguistic needs of the respective communicative situations. For instance, *also* occurs more often as an adverbial consecutive connector in written productions, but as a repair marker in spoken ones; the function of an elaboration

[13] For a detailed analysis of openings and closings, based on findings from all five languages, see Katsika et al. (2025 [this volume]).

or specification marker is more frequent in the formal conditions, whereas it is used for evaluation more often in the informal ones.

A similar register-differentiated distribution can be found for a graphic element, namely the three-dot sign ⟨…⟩. As a graphic element, the three-dot sign is restricted to written productions. Within this mode, it shows clear functional distinctions along the parameter of (in)formality: in productions from the formal-written condition, it is used exclusively as a placeholder for lexical material, in those from the informal-written condition, it is used in discourse-pragmatic functions, for instance as a discourse-organisational device (see Labrenz et al. (2025 [this volume]) and Labrenz et al. (2022) for more details).

Another kind of graphic element we found in our data are emoji. Emoji have been shown to function as markers of informality (Siebenhaar 2020), and in line with this, in our corpus they occur exclusively in the WhatsApp messages, that is, in the informal-written condition, where they can be analysed as graphic discourse markers (Wiese & Labrenz 2021 for details).

While the three-dot sign and emoji are translinguistic elements by virtue of their graphic nature, we also found evidence for cross-linguistic lexical integration, and again, this reflected register-differentiations. Since our elicitation procedure favoured a monolingual mode (see Section 2.2 above), we did not expect to find a lot of language mixing. However, while not very frequent, we did find some evidence for this. ((9)) through ((16)) illustrate this with data from the German (9 through 11), English (12), Russian (13), Greek (14), and Turkish ((15), (16)) subcorpora.

The phenomena include code-switching between majority and heritage language (9), the use of discourse markers and interjections from international English outside the US ((9), (10)), from the heritage language in the majority language ((11), (12)), and vice versa ((13) through (16)).

(9) und **guess what** in dem Moment kamen zwei Autos
and guess what$_{\text{English}}$ in that moment came two cars

hintereinander und ja **kopek sokagin ortasina gidince**
behind.each.other and yes dog street in.the.middle when.goes

araba ani fren yapti
car suddenly brake make$_{\text{Turkish}}$

'And *guess what*, at that moment two cars came up behind each other and yes, *the car breakes suddenly as the dog runs into the middle of the road.*'

[DEbi01FT_iwD]

(10) Alter, ich bin gerade so nach Hause gelaufen und da ist ja
old.one, I have just like to home walked and there is yes
dieser Parkplatz, **you know**?
this car.park *you know*$_{\text{English}}$?
'Dude, I was just walking home and there's this car park, *you know*?'
[DEmo53FD_iwD]

(11) Die riefen danach die Polizei. **Hadi** bis spaeter dicker
they called afterwards the police. *OK then*$_{\text{Turkish}}$ until later fat.one
'They called the police afterwards. *OK then*, see you later dude'
[DEbi04MT_iwD]

(12) *Okay* **privet** [name], so I just saw this car accident and it was kind of weird
'Okay *hi*$_{\text{Russian}}$ [name], so I just saw this car accident and it was kind of weird' [USbi67MR_isE]

(13) первая машина так затормозила что вторая **цак** и
first car so braked that second *wham*$_{\text{German}}$ and
въехала!
drove.in
'the first car braked so hard that the second *wham* and drove into it!'
[DEbi18MR_iwR]

(14) αυτά οκέι τα λέμε **τσους**
that.was.it okay these talk *bye*$_{\text{German}}$
'that was it okay see you then *bye*' [DEbi07FG_isG]

(15) **omg** kız bugün araba kazasını gördüm !
omg$_{\text{English}}$ girl today car accident saw
'*omg* girl I saw a car accident today!' [USbi02FT_iwT]

(16) nerdeyse araba da hemen e **vollbremsen** gibi
almost car also immediately uh *fully.break*$_{\text{German}}$ like
bir şey yaptı **einfach** hemen durdu arkadaki
something did *simply*$_{\text{German}}$ immediately stopped in.the.back
araba da arabada da [laughing] öndekine çarptı
car also in.the.car also *[laughing]* the.one.in.front hit

pek bir şey olmadı **eigentlich**
not.much something happened *after.all*$_{\text{German}}$
'almost immediately the car did something like a *full stop* and *simply* stopped immediately and the car behind hit the car in front, nothing much happened *after all*' [DEbi52FT_isT]

Such patterns were limited to productions from the informal conditions.[14] This is what would be expected for naturalistic register differentiations: in the countries we investigated, monoglossic language ideologies generally stigmatise language mixing in formal settings, and translinguistic fluidity and creativity is mostly restricted to informal contexts with friends. Hence, cross-linguistic integration can serve as an informality marker. What is interesting is that we can observe some evidence for this even though our elicitations were conducted in a generally monolingual mode. This, then, further underlines the power of our set-up to elicit register distinctions that are close to those expected in spontaneous data. The following section targets this aspect through comparison with spontaneous data.

3.2 Parallels with spontaneous data

3.2.1 Informal spontaneous productions as a point of comparison

For our comparison, we focussed on informal-spoken data since this is the kind of data that is notoriously difficult to elicit. A previous study revealed similarities of LangSit data and spontaneous data for the example of bare local NPs. In this study, Wiese & Pohle (2016) compared informal-spoken productions of adolescents from a smaller corpus of LangSit data with spontaneous spoken productions from the KiDKo corpus:[15] this corpus is based on self-recordings conducted by adolescents in peer-group situations. In both kinds of data, they found bare local NPs that would be ungrammatical in standard German, and these NPs followed the same syntactic and semantic patterns in the elicited LangSit data and the spontaneous Kiezdeutsch data. This is a first indication that the LangSit method is suitable to elicit naturalistic data, that is, language productions that are similar to spontaneous ones.

[14] In majority language use, cross-linguistic integration was generally restricted to informal productions. In heritage language use, there were also some integrations in formal productions, but these were lexical elements like nouns and verbs rather than discourse markers (and there was no code switching).

[15] Wiese et al. 2012; https://hu.berlin/kidko (last accessed: 2024-10-23).

In the present study, we followed up on these results with a broader comparison based on the informal-spoken LangSit data from the RUEG corpus and naturally occurring spontaneous data of a kind that was maximally close to the RUEG data.

The corpus of spontaneous data (the "Roy corpus"; see Roy 2022) consists of 15 not-elicited voice messages sent to friends via WhatsApp that had been donated to one of us (A. Roy), who knew all donors personally. This data set is part of a larger set of donations; the criterion for inclusion in the corpus was that the messages needed to contain some narration, given that the RUEG data was based on narrations of the accident. The speakers were young adults in Germany who had all grown up in families with German as the only family language.

To generate a comparative data set from the RUEG corpus, we extracted 15 informal-spoken productions from the German subcorpus, of the speaker group of monolingual adults in Germany, choosing speakers who were similar to the donors with respect to age/generation and education: speakers' average year of birth was 1991 in the Roy corpus and 1992 in the RUEG subcorpus, their average age was 28 and 27 years, respectively, and the majority of speakers had a university degree (11 and 8 of the speakers, respectively). The spontaneous data was transcribed in the same way as the RUEG data. For our analyses, we mostly concentrated on the narrative parts (except for the investigation of discourse openings), that is, we ignored organisational passages that often occurred at the beginning or end in the spontaneous spoken data. Counting only the narratives from the spontaneous data, this yielded a data set with a size of 3,450 words and 327 communicative units (CUs, defined as higher-level segmentation units consisting of an "independent clause with its modifiers", Loban 1976: 9). By comparison, the size of the RUEG extract was 2,931 words and 298 CUs.

For our investigation, we targeted all noncanonical patterns, which yielded 30 types of phenomena. Analyses revealed far-reaching similarities of the patterns in both data sets. In what follows, we describe some central findings at levels of syntax, discourse organisation, and lexicon; a detailed discussion of all results can be found in Roy (2022). Given the small size of the comparative (sub)corpora, we focus on qualitative patterns.

3.2.2 Syntax

Syntactic parallels were evident in a number of word order patterns that are characteristic for informal-spoken German and would not be expected in formal and/or written registers that are usually close to the standard. Examples are

verb-first (V1) declaratives, right dislocations of modal expressions, and left dislocations in a topic-comment structure (see, for instance, Önnerfors 1997, Imo 2014 for such patterns in German). (17) through (19) give examples from both data sets. Note that the V1 examples also include ones that do not contain an ellipsis of obligatory constituents ((17b) through (17d)). This is a type considered "genuine V1" in the literature (e.g. Önnerfors 1997), which are generally less acceptable (Auer 1993) and which we found in both data sets.

(17) V1

a. **hatte** einen zahlendreher drin **hab** mich zum dritten mal
had a reversal.of.numbers in **have** me for.the third time
mit meinem text vorgestellt
with my text indroduced
'had a reversal of numbers, introduced myself with my text for the third time' [Roy, Bmo9fD_1]

b. und **is** ihr halber Einkauf auf die Straße gefallen (ne)
and **is** her half purchase on the road fallen QUESTION.TAG
'and half her purchase has fallen on the road, right'
[RUEG, DEmo43FD_isD]

c. und **sagt** die Person so ja sie sind hier gar nicht bei
and **says** the person like yes you$_{[V\text{-}FORM]}$ are here at.all not at
dem Amt
the department
'and the person says, like, well, you aren't at the department here at all' [Roy, Bmo9fD_1]

d. ä:h genau **muss** ich auch noch warten
uh exactly **must** I also still wait
'uh right, on top of everything I need to wait'
[RUEG, DEmo39FD_isD]

(18) Right dislocation

a. andererseits als ich dann vorgelesen habe habe ich mir auch
on.the.other.hand when I then read.out have have I me also
so ein bisschen bewusst gemacht was der Inhalt eigentlich ist
such a little.bit aware made what the content actually is
irgendwie
somehow
'on the other hand when I read it out loud, I somehow became a little bit aware of what the content was, actually' [Roy, Bmo2fD_1]

b. und du glaubs es nich wat [laughing] was ich hier was äh
and you believe it not what [laughing] what I here what uh
gesehen hab **schon wieder**
seen have **yet again**
'and you won't believe what uh I saw yet again'
[RUEG, DEmo43MD_isD]

(19) Left dislocation

a. ja aber **dieses system so was so den Stoff aufzubauen das**
yes but **this system** like that like the stuff to.build.up **that**
kommt irgendwie schwer in meinen Kopf
comes somehow hard in my head
'yes but this system how to build up the stuff, I somehow don't get it'
[Roy, Bmo18mD_1]

b. **der Mann der** hat die ganze Zeit mit =n Ball mit m dem
the man who/he has the whole time with =a ball with a
Fußball gedribbelt
football dribbled
'the man, he dribbled with a ball – with a football all the time'
[RUEG, DEmo17MD_isD]

Of the V1-sentences that can be analysed as pronoun ellipses, we found differences with respect to the category of person: ellipsis was mostly of 1st person singular pronouns in the spontaneous data, and mostly of 3rd person singular pronouns in the elicited data. This is presumably due to the difference in the topic under discussion, which in the RUEG data was always an event where the speaker was not involved. Note that despite these quantitative differences, we found both patterns in both data sets.

3.2.3 Discourse organisation and lexicon

When we look at openings, we find similar patterns in both data sets, namely those prototypical for narratives, where speakers use an attention-getter and make the narrative relevant (e.g. Quasthoff 1979, Ochs & Capps 2002: Chapter 1, Gülich 2020 for German; see examples in Section 3.1.2 from the RUEG corpus). Not all messages include such an opening, though, and the frequency is higher in the RUEG data: all voice messages in the RUEG subcorpus start with an opening, compared to only 2 of the spontaneous messages. This can be attributed to the context in which the respective messages occur. As mentioned in Section 3.1.2

above, the RUEG narratives were not embedded in a larger conversation. In contrast, the spontaneous messages were often integrated. Accordingly, in the Roy corpus, speakers would presumably have felt less of a need to make their narrative relevant and could assume tellability without any comment (c.f. König 2019). (20) gives an example from the spontaneous data set where this is particular clear. In this message, the speaker talks about the behaviour of a mutual friend and starts her narrative without any introduction, beginning the message with this:

(20) äh na ja also sie war irgendwie so voll überschwänglich
uh well yes so she was somehow like really effusive
'uh well, so somehow she was like really effusive' [Roy, Bmo5fD_4]

Both data sets alike contained a range of discourse phenomena typical of narratives, for instance markers of moral stance where speakers evaluate the narrative or a part of it (21 and 22), references to common ground ((23) and (24)), and hesitations at points of complication in the narrative ((25) and (26)).

(21) es war nett aber auch irgendwie komisch
it was nice but also somehow strange
'it was nice but also strange somehow' [evaluating a university seminar]
[Roy, Bmo3fD_1]

(22) es war ein bisschen unsinnig
it was a little.bit nonsensical
'it was a little bit nonsensical' [evaluating the accident]
[RUEG, DEmo18FD_isD]

(23) genau jetzt bin ich grad mit [name] und [name] auf
exactly now am I just with [name] and [name] on
[name of an island] ist total geil
[name of an island] is totally cool
'right, I am with [name] and [name] on [name of an island] now and it's totally cool' [Roy, Bmo13mD_1]

(24) den Parkplatz an sich den kennste ja da is ja die Hauptstraße
the car.park at itself that knows.you yes there is yes the main.road
vorne
in.front
'you know the car park after all, there is this main road in front of it'
[RUEG, DEmo17MD_isD]

(25) allerdings beim Aufwachen äh hatte er Lungenprobleme und
although during.the waking.up uh had he lung.problems and
somit auch äh wohl schwerwiegende Atemprobleme äh
accordingly also uh probably serious breathing.problems uh
sodass er dann noch ne Nacht äh auf der Intensivstation bleiben
so.that he then still a night uh on the intensive.care.unit stay
musste
had.to
'but waking up he had problems with the lungs and accordingly probably also serious breathing problems and so he needed to stay in the intensive care unit for another night' [narrative about an operation that went well but then some complications occured afterwards] [Roy, Bmo1fD_1]

(26) und äh dann war da so n Hund der äh der halt auf den Ball
and uh then was there such a dog that uh that just on the ball
abgegangen ist
on.went has
'and uh then there was this dog that uh just went for the ball'
[RUEG, DEmo18FD_isD]

Again, we found the same patterns in both kinds of data. At the lexical level, we found similar patterns for instance for discourse markers in the left periphery. Qualitative analysis indicated that such discourse markers occurred at the same narrative loci in both data sets. Typical loci were, e.g., at the beginning of a narrative (cf. (27) and (28)) or of a longer explanation ((29) and (30)):

(27) äh **also** die OP von [name] ist gut verlaufen
uh **so** the OP(eration) of [name] has well gone
'uh **so** the operation of [name] went well' [Roy, Bmo1fD_1]

(28) **und zwar** ähm wir ham uns ja grad verabschiedet
and **in.fact** uh we have us yes just said.goodbye
'and uh we just said goodbye to each other' [RUEG, DEmo17MD_isD]

(29) **also** ich wollte es kaufen damit er seine Zeit nicht vergeudet
so I wanted it buy so.that he his time not wastes
'So I wanted to buy it so he doesn't waste his time' [Roy, Bmo8mD_1]

(30) **also** der konnte gar nich groß auf die Straße renn
so he could at.all not big on the road run
'so he couldn't really run onto the road properly'
[RUEG, DEmo20FD_isD]

Taken together, we hence find parallels of the RUEG corpus data with spontaneous data at levels of syntax, discourse organisation, and lexicon, and we find them for informal-spoken productions, which represent a register that is particularly difficult to elicit, given its comparably large distance from standard language.

4 Conclusions and outlook

In this paper, we addressed the methodological challenges one faces when aiming for data that is ecologically valid and representative of speakers' repertoires. We argued that such data is important if we want to account for actual linguistic patterns and linguistic variation. In particular we argued that we cannot confine ourselves to data elicited under conditions that favour formal language use in, e.g., a lab setting. To get a realistic picture of linguistic practices and language competence, we need to cast our web wider and include informal as well as formal communicative situations, ideally across both written and spoken modes. Including a broader range of registers also allows us to tease apart phenomena associated with informal registers and nonstandard patterns that might be due to other factors, for instance language contact, heritage language developments, or acquisitional stages in language learning. For an informed analysis of linguistic patterns, we hence need to use ecologically valid data from a range of communicative settings, and we need to do so for multilingual and monolingual speakers alike.

This is what the RUEG group set out to do, and in this paper, we described and discussed the methodology we used for creating a database for this endeavour. Since our shared enterprise included large-scale comparisons, our data needed to be controlled enough to allow for comparisons across different communicative situations, languages, speaker groups, and (contact-linguistic) contexts. Accordingly, we needed to conduct controlled elicitations. For this, we developed a method that was based on the "Language Situations" set-up where participants see a nonverbal (video or photo) stimulus showing an interesting event, for instance, a minor traffic accident, and are asked to imagine witnessing this incident and play-act telling different interlocutors about it (Wiese 2020).

We described the specifications we made to this general set-up in order to ensure homogeneity across our large-scale elicitations and to maximally support naturalistic language productions. As we have shown in this paper, our elicitation method was successful in yielding data that was (a) representative of different communicative situations, and (b) close to naturally occurring data.

First, we showed that the data we elicited in the formal vs. informal, and written vs. spoken conditions showed systematic differences that indicate register-differentiated productions. We discussed examples for such differences at levels of syntax, discourse organisation, and the lexicon. Our analyses revealed qualitative and quantitative differences along the lines expected for register distinctions.

Second, we showed that the corpus data displays similar patterns as spontaneously occurring data. We targeted informal-spoken productions because this is a type of data that is notoriously difficult to elicit. We compared RUEG corpus data of German monolingual adults with spontaneous messages by a comparable speaker group. Again, we discussed examples from syntax, discourse, and the lexicon, and showed that we find the same patterns in both data sets.

Taken together, this underlines the value of our methodology for collecting ecologically valid, register-differentiated data. It makes the RUEG corpus a suitable basis for investigations that do not just evaluate noncanonical patterns against the yardstick of a putative monolingual standard language: using the RUEG corpus, we can systematically tap into speakers' repertoires, taking into account both formal and informal registers, and doing so for multilinguals and monolinguals alike.

Our discussion pointed to a number of aspects one should keep in mind when implementing this methodology. We see three major points. First, a frequent challenge for data gathering are the resources available, in terms of time, money, and personnel. The RUEG group elicited data in 5 countries, for 5 languages, in a large range of different contact settings, from 2 age groups, and from bi- and monolingually raised speakers. This makes our corpus a rich database with a huge potential for further research (see also Shadrova et al. (2025 [this volume]) on the RUEG corpus): it provides a basis for a wealth of different comparisons that we hope will continue to be a useful resource for other researchers as well. Creating such a database was possible because of the large number of projects and researchers our group has brought together. However, applying the LangSit method is also possible with much less resources. As we argued in this paper, one major advantage is that it is less time- and cost-intensive than gathering spontaneous, naturally occurring data, but at the same time yields data that is naturalistic in the sense of showing parallels to such spontaneous data. Applying the LangSit method to a smaller number of participants, e.g., concentrating on a

specific speaker group, does not require a large amount of resources, and it can be done even in the context of a Master thesis (see Wiese 2020 for examples of this). Such smaller data sets can also be used for larger comparisons with existing data from the RUEG corpus, in particular if they were elicited with the same stimuli (which is straightforward since all RUEG resources including stimuli are open access, as described in our discussion).

A second point are the requirements on elicitation. In order to support participants' natural productions, it is important to create a relaxed atmosphere. We have argued that this is particularly true for the informal productions, since language in informal communicative situations often deviates from standard language, is thus considered "incorrect" and might be modified in a situation perceived as testing. As described, setting up an informal atmosphere to avoid this can be achieved through room decoration, elicitor's habitus (clothes, behaviour) and introductory chit-chat sessions. Nevertheless, one should expect at least some participants to have problems (play-)acting naturally, and take this into account in analysis.

A third aspect to keep in mind – kind of the other side of the coin – is the degree to which the data is controlled. On the one hand, as we discussed in this paper, the choice of the event presented in the stimulus guides speakers towards specific topics, and this is something that analyses should take into account (for instance, the topic under discussion can have an impact on the frequency of nouns vs. pronouns). On the other hand, LangSit data is, by virtue of its naturalistic nature, not as controlled as experimental data, and this means that for our analyses we face a challenge that is well known from corpus linguistics in general: some phenomena might not occur, and their absence does not prove that they are not part of a speaker's repertoire or that they are ungrammatical in a variety or a language. Hence, LangSit data is not suitable for providing negative evidence. As we have shown in this paper, its strengths lie in revealing novel and unexpected grammatical and pragmatic patterns in naturalistic language productions, in capturing a broader part of speakers' linguistic repertoires, and in allowing meaningful comparisons of bilingual and monolingual data. As such, LangSit data can reveal interesting linguistic patterns in language use and provide evidence for parallels and differences between speaker groups and (contact-linguistic) settings. If negative evidence is needed for further testing the details of such patterns, LangSit findings can be complemented by experimental data, e.g., acceptability judgments or sentence completion tasks. As several of the RUEG projects have shown, this kind of multi-method approach can be particularly fruitful for investigations into morphological and syntactic structure (see Tsehaye et al. (2025 [this volume]) and Özsoy et al. (2025 [this volume]) for examples).

We see set-ups of such controlled experiments as another domain that might benefit from the results of the methodological discussion we presented in this paper. In particular, we should aim for such experiments to also allow for register distinctions, and to do this in a way that encourages naturalistic language use. For instance, an experiment might not only be conducted in a formal lab room, but also in more informal surroundings, and stimuli should account for informal as well as formal communicative situations, e.g., through target sentences in acceptability tests that are introduced as examples from either a WhatsApp message to a friend or from a formal report. This way, our findings might inspire further set-ups for controlled experiments, with a potential for revealing novel findings on monolinguals and bilinguals alike.

Finally, as we described in this paper, our elicitations favoured a monolingual mode in both of heritage speakers' languages, given that in a first step, we wanted to target developments within their heritage and their majority language varieties. In future research, one should also include bilingual modes that encourage language mixing, especially in informal settings. Such mixing is something that is common even in societies dominated by a monolingual habitus (as is the case in the countries we investigated), and it showed up in our data even though the set-up did not support this. This points to the normality of such mixing in natural language use and should encourage us to further move away from bilingual/-monolingual dichotomies and take a more translinguistic perspective on grammatical patterns and linguistic resources. The methods we discussed in this paper lend themselves to such research, and we hope they will be a useful resource for future investigations into native grammars and linguistic diversity, taking into account the breadth of speakers' linguistic repertoires and their actual language use in different communicative situations.

Acknowledgements

We would like to thank Maria Pohle and our fantastic and dedicated student assistants Luisa Koch, Sabine Hainsfurth, Claudia Czarniak, and Tjona Sommer. Maria Pohle was involved in the first two years of project Pd-2 and was very much missed in the subsequent years. Her contribution to the project is gratefully acknowledged. We also thank the anonymous reviewers for their helpful and very constructive suggestions, and Lea Coy for help with pre-publication formatting. The research was supported through funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the Research Unit *Emerging Grammars in Language Contact Situations*, project Pd (grant number: 394844736).

References

Aalberse, Suzanne P., Ad Backus & Pieter Muysken (eds.). 2019. *Heritage languages: A language contact approach* (Studies in Bilingualism 58). Amsterdam: John Benjamins. DOI: 10.1075/sibil.58.

Adli, Aria & Gregory R. Guy. 2022. Globalising the study of language variation and change: A manifesto on cross-cultural sociolinguistics. *Language and Linguistics Compass* 16(5-6). DOI: 10.1111/lnc3.12452.

Auer, Peter. 1993. Zur Verbspitzenstellung im Gesprochenen Deutsch. *Deutsche Sprache* 23. 193–222.

Bayram, Fatih, Tanja Kupisch, Diego Pascual Y Cabo & Jason Rothman. 2019. Terminology matters on theoretical grounds, too! Coherent grammars cannot be incomplete. *Studies in Second Language Acquisition* 41(2). 257–264. DOI: 10.1017/S0272263119000287.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas & Susan Conrad. 2019. Analyzing linguistic features and their functions. In *Register, genre, and style*, 2nd edn., 51–84. Cambridge: Cambridge University Press. DOI: 10.1017/9781108686136.

Bilá, Magdaléna, Alena Kačmárová & Ingrida Vaňková. 2020. The encounter of two cultural identities: The case of social deixis. *Russian Journal of Linguistics* 24(2). 344–365. DOI: 10.22363/2687-0088-2020-24-2-344-365.

Bunk, Oliver, Shanley E. M. Allen, Sabine Zerbian, Tatiana Pashkova, Yulia Zuban & Erica Conti. 2025. Information packaging and word order dynamics in language contact. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 379–412. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775181.

Granger, Sylviane. 2008. Learner corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 259–275. Boston: De Gruyter. DOI: 10.1515/9783110213881.2.

Gülich, Elisabeth. 2020. Alltägliches erzählen und alltägliches Erzählen. In Elisabeth Gülich, Stefan Pfänder, Carl Eduard Scheidt & Elke Schumann (eds.), *Mündliches Erzählen: Verfahren narrativer Rekonstruktion im Gespräch* (Narratologia), 3–26. Boston: De Gruyter.

Imo, Wolfgang. 2014. Elliptical structures as dialogical resourses for the management of understanding. In Susanne Günthner, Wolfgang Imo & Jörg Bücker (eds.), *Grammar and dialogism: Sequential, syntactic, and prosodic patterns be-*

tween emergence and sedimentation (Linguistik: Impulse & Tendenzen 61), 139–176. Boston: De Gruyter.

Katsika, Kalliopi, Annika Labrenz, Kateryna Iefremenko & Shanley E.M. Allen. 2025. Discourse openings and closings across languages in contact. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 527–576. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775189.

Klotz, Martin, Rahel Gajaneh Hartz, Annika Labrenz, Anke Lüdeling & Anna Shadrova. 2024. Die RUEG-Korpora: Ein Blick auf Design, Aufbau, Infrastruktur und Nachnutzung multilingualer Forschungsdaten. *Zeitschrift für germanistische Linguistik* 52(3). 578–592. DOI: 10.1515/zgl-2024-2026.

König, Katharina. 2019. Narratives 2.0: A multi-dimensional approach to semi-public storytelling in WhatsApp voice messages. *Journal für Medienlinguistik* 2(2). 30–59. DOI: 10.21248/jfml.2019.10.

Kupisch, Tanja & Maria Polinsky. 2022. Language history on fast forward: Innovations in heritage languages and diachronic change. *Bilingualism: Language and Cognition* 25(1). 1–12. DOI: 10.1017/S1366728921000997.

Kupisch, Tanja & Jason Rothman. 2018. Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism* 22(5). 564–582. DOI: 10.1177/1367006916654355. (27 April, 2023).

Labrenz, Annika. 2023. Functional variation of German *also* across registers and speaker groups. *Contrastive Pragmatics* 4(2). 289–320. DOI: 10.1163/26660393-bja10077.

Labrenz, Annika, Kateryna Iefremenko, Kalliopi Katsika, Shanley E.M. Allen, Christoph Schroeder & Heike Wiese. 2025. Dynamics of discourse markers in language contact. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 493–526. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775187.

Labrenz, Annika, Heike Wiese, Tatiana Pashkova & Shanley E. M. Allen. 2022. The three-dot sign in language contact. *Pragmatics & Cognition* 29(2). 246–271. DOI: 10.1075/pc.21021.lab.

Loban, Walter. 1976. *Language development: Kindergarten through grade twelve*. Tech. rep. 18. National Council of Teachers of English, 1111 Kenyon Road, Urbana, Illinois 61801 (Stock No. 26545). https://eric.ed.gov/?id=ED128818 (27 April, 2023).

Lüdeling, Anke, Artemis Alexiadou, Shanley E. M. Allen, Oliver Bunk, Natalia Gagarina, Sofia Grigoriadou, Rahel Gajaneh Hartz, Kateryna Iefremenko, Esther Jahns, Kalliopi Katsika, Mareike Keller, Martin Klotz, Thomas Krause, Annika Labrenz, Maria Martynova, Onur Özsoy, Tatiana Pashkova, Maria Pohle, Judith Purkarthofer, Vicky Rizou, Christoph Schroeder, Anna Shadrova, Luka Szucsich, Rosemarie Tracy, Wintai Tsehaye, Heike Wiese, Sabine Zerbian, Yulia Zuban & Nadine Zürn. 2024. *RUEG Corpus*. Version 1.0. Zenodo. DOI: 10.5281/zenodo.11234583.

Montrul, Silvina. 2015. *The acquisition of heritage languages*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139030502.

Ochs, Elinor & Lisa Capps. 2002. *Living narrative: Creating lives in everyday storytelling*. Cambridge: Harvard University Press.

Önnerfors, Olaf. 1997. *Verb-erst-Deklarativsätze: Grammatik und Pragmatik* (Lunder germanistische Forschungen 60). Stockholm: Almqvist & Wiksell International.

Özsoy, Onur, Vasiliki Rizou, Maria Martynova, Natalia Gagarina, Luka Szucsich & Artemis Alexiadou. 2025. Null subjects in heritage Greek, Russian and Turkish. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 179–217. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775167.

Pashkova, Tatiana, Marlene Böttcher, Kalliopi Katsika, Sabine Zerbian & Shanley E. M. Allen. 2025. Majority English of heritage speakers. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 449–491. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775185.

Polinsky, Maria. 2018. *Heritage languages and their speakers* (Cambridge Studies in Linguistics 159). Cambridge: Cambridge University Press. DOI: 10.1017/9781107252349.

Quasthoff, Uta. 1979. Die Partikeln der deutschen Sprache. In Harald Weydt (ed.), *Verzögerungsphänomene, Verknüpfungs- und Gliederungssignale in Alltagsargumentationen und Alltagserzählungen*, 39–57. Boston: De Gruyter. DOI: 10.1515/9783110863574.39.

Rothman, Jason, Fatih Bayram, Vincent DeLuca, Grazia Di Pisa, Jon Andoni Duñabeitia, Khadij Gharibi, Jiuzhou Hao, Nadine Kolb, Maki Kubota, Tanja Kupisch, Tim Laméris, Alicia Luque, Brechje van Osch, Sergio Miguel Pereira Soares, Yanina Prystauka, Deniz Tat, Aleksandra Tomić, Toms Voits & Stefanie Wulff. 2023. Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics* 44(3). 316–329. DOI: 10.1017/S0142716422000315.

Roy, Albrun. 2022. *Elizitierte Informalität: Ein Vergleich elizitierter und spontansprachlicher Sprachnachrichten.* Supervisor: Heike Wiese. Universität Potsdam. (MA thesis).

Schroeder, Christoph, Kateryna Iefremenko & Mehmet Öncü. 2024. The postverbal position in heritage Turkish: A comparative perspective with a focus on non-clausal elements. In Zeynep Kalkavan-Aydın & Yazgül Şimşek (eds.), *Deutsch-Türkische Zweisprachigkeit mit besonderem Fokus auf Jugendliche [Turkish-German bilingualism with a special focus on adolescents]* (Mehrsprachigkeit [Multilingualism] 58), 109–132. Münster: Waxmann.

Sevinç, Yeşim & Jean-Marc Dewaele. 2018. Heritage language anxiety and majority language anxiety among Turkish immigrants in the Netherlands. *International Journal of Bilingualism* 22.2. 159–179. DOI: 10.1177/1367006916661635.

Shadrova, Anna, Martin Klotz, Rahel Gajaneh Hartz & Anke Lüdeling. 2025. Mapping the mappings and then containing them all: Quality assurance, interface modeling, and epistemology in complex corpus projects. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 69–110. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775161.

Siebenhaar, Beat. 2020. Informalitätsmarkierung in der WhatsApp-Kommunikation. In Jannis Androutsopoulos & Florian Busch (eds.), *Register des Graphischen: Variation, Interaktion und Reflexion in der digitalen Schriftlichkeit*, 67–92. Boston: De Gruyter. DOI: 10.1515/9783110673241-004.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Tsehaye, Wintai, Rosemarie Tracy & Johanna Tausch. 2025. Inter- and intra-individual variation: How it materializes in Heritage German and why it matters. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 141–177. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775165.

Wiese, Heike. 2013. What can new urban dialects tell us about internal language dynamics? The power of language diversity. In Werner Abraham & Elisabeth Leiss (eds.), *Dialektologie in neuem Gewand: Zu Mikro-/Varietätenlinguistik, Sprachenvergleich und Universalgrammatik* (Linguistische Berichte Sonderheft 19), 207–245. Hamburg: Buske.

Wiese, Heike. 2020. Language situations: A method for capturing variation within speakers' repertoires. In Yoshiyuki Asahi (ed.), *Methods in dialectology XVI* (Bamberg Studies in English Linguistics 59), 105–117. Frankfurt am Main: Peter Lang.

Wiese, Heike, Artemis Alexiadou, Shanley E. M. Allen, Oliver Bunk, Natalia Gagarina, Katerina Iefremenko, Maria Martynova, Tatiana Pashkova, Vicky Rizou, Christoph Schroeder, Anna Shadrova, Luka Szucsich, Rosemarie Tracy, Wintai Tsehaye, Sabine Zerbian & Yulia Zuban. 2022. Heritage speakers as part of the native language continuum. *Frontiers in Psychology* 12. 5982. DOI: 10.3389/fpsyg.2021.717973.

Wiese, Heike, Shanley E. M. Allen, Mareike Keller & Artemis Alexiadou. 2025. Introduction: Investigating the dynamics of language contact situations. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Linguistic dynamics in heritage speakers: Insights from the RUEG group*, 1–29. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775157.

Wiese, Heike, Ulrike Freywald, Sören Schalowski & Katharina Mayr. 2012. Das KiezDeutsch-Korpus: Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache* 2. 97–123. DOI: 10.37307/j.1868-775X.2012.02.02.

Wiese, Heike & Annika Labrenz. 2021. Emoji as graphic discourse markers: Functional and positional associations in German WhatsApp messages. In Daniël Van Olmen & Jolanta Šinkūnienė (eds.), *Pragmatic markers and peripheries* (Pragmatics & Beyond New Series), 277–302. Amsterdam: John Benjamins. DOI: 10.1075/pbns.325.10wie.

Wiese, Heike & Hans Georg Müller. 2018. The hidden life of V3: An overlooked word order variant on verb-second. In Mailin Antomo & Sonja Müller (eds.), *Non-canonical verb positioning in main clauses* (Linguistische Berichte Sonderheft 25), 202–223. Hamburg: Buske.

Wiese, Heike & Maria Pohle. 2016. „Ich geh Kino" oder „… ins Kino"? Gebrauchsrestriktionen nichtkanonischer Lokalangaben. *Zeitschrift für Sprachwissenschaft* 35(2). 171–216. DOI: 10.1515/zfs-2016-0012.

Wiese, Heike, Horst Simon, Christian Zimmer & Kathleen Schumann. 2017. German in Namibia: A vital speech community and its multilingual dynamics. Péter Maitz & Craig A. Volker (eds.). 221–245.

Zimmer, Christian, Heike Wiese, Horst J. Simon, Marianne Zappen-Thomson, Yannic Bracke, Britta Stuhl & Thomas Schmidt. 2020. Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. *Deutsche Sprache* 48(3). 210–232. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-100570 (27 April, 2023).