



Volume
19 (2025)

Pages
40 - 66

Received
27 Nov 2024

Handle
<https://hdl.handle.net/10125/74810>

Online
<http://nflrc.hawaii.edu/ldc>

Citation
Lehmann, Nico, Vahid Morteza pour, Jozina Vander Klok, Zahra Farokhnejad, David Müller, Elisabeth Verhoeven, Aria Adli. 2025. Lang*Reg corpus: Documenting intra-speaker variation across languages and registers. *Language Documentation & Conservation* 19: 40-66.

Lang*Reg corpus: Documenting intra-speaker variation across languages and registers

Nico Lehmann
Humboldt-Universität zu Berlin

Vahid Morteza pour
Universität zu Köln

Jozina Vander Klok
Humboldt-Universität zu Berlin

Zahra Farokhnejad
Universität zu Köln

David Müller
University of Geneva

Elisabeth Verhoeven
Humboldt-Universität zu Berlin

Aria Adli
Universität zu Köln



Language Documentation & Conservation

We present a new corpus design for multi-lingual corpora that involve intra-speaker variation in different situational-functional contexts, including primarily spoken but also the written mode, with the aim towards enhancing language documentation efforts and resources. We illustrate how this comparative design and the resulting cross-culturally applicable data collection procedure has been successfully realized in order to build the Lang*Reg corpus (Adli et. al. 2024), which currently includes five languages from three different language families: German, Persian, Southern Kurdish, Yucatec Maya and Javanese. For each of these languages, the same native speakers were asked to produce language in two types of activities that naturally occur in all the respective cultural contexts: telling a story to a friend, and talking freely with various interlocutors (friend, stranger, taxi driver, university professor). Moreover, our design included the storytelling in two modes, which allows for the comparison between spoken and written modes of the same language user. We show how Lang*Reg provides a versatile resource for many purposes – in particular research into register due to the variety of situational contexts involved, we show how German and Persian exploit the right periphery for different register distinctions, and we invite others to use this resource. At the same time, we show how the methodology developed can be used as a template to complement language resources by creating comparable intra-individual, multi-purpose data sets.

1. Introduction¹

Although an increasing amount of language data is collected and published in corpora, two characteristics are still rare in current data collections: few data sets are compiled with equivalent collection methods for different languages, thus making them as comparative as possible across languages, and even fewer resources have incorporated intra-individual variation systematically. With the Lang*Reg corpus (Adli et al. 2024), we set out to provide an exemplary corpus incorporating intra-individual variation across multiple situations, including primarily the spoken but also the written mode, in five languages across three language families – German [deu] (Indo-European, West-Germanic); Persian [per] and Southern Kurdish [sdh] (Indo-European, Indo-Iranian); Yucatec Maya [yua] (Mayan); and Javanese [jav] (Austro-nesian). The language sample has been selected in order to investigate the role of cross-linguistic and cross-cultural aspects in register variation. Hence, we included languages which differ along pertinent dimensions including typological, socio-cultural and ecological properties to be detailed in §5.

The main goal of this paper is to introduce the Lang*Reg corpus for the purposes of (i) illustrating how one may create comparable cross-linguistic data when including the same speakers across several production circumstances, and (ii) inviting scholars to use Lang*Reg as a data resource in linguistic analysis or as a template for replication in language documentation and beyond.

The intra-individual design is in particular relevant for register analyses, which is intra-individual variation in linguistic behaviour resulting from aspects of the situational-functional setting (Biber 2012; Biber & Conrad 2019; Lüdeling et al. 2022) because linguistic phenomena “serve the communicative functions required by the situational contexts of the texts” (Biber et al. 2021: 22). With Lang*Reg we can directly observe how speakers adapt their language output in different situational-functional contexts and compare such behaviours across languages. Most existing corpora only allow for indirect observations of register variation because they do not involve the same speakers in a controlled setup or are limited as to the contexts included (see §2). Early results using Lang*Reg show different register-dependent effects on pronoun rates in distinct types of argument drop languages: in the pro-drop language Persian, (c)overt subject pronoun rates are similar across registers whereas in the topic drop language German, (c)overt subject pronoun rates are higher in more informal registers. With respect to syntactic complexity, preliminary Lang*Reg results substantiate previous findings that syntactic complexity varies robustly across languages by a number of register-related parameters, in particular mode (Biber 1988; Biber & Gray 2010) and distance vs. closeness (Verhoeven & Lehmann 2018). Word order phenomena are also well-suited for investigating register variation: Lang*Reg indicates, for instance, that certain right-peripheral subjects, which are more tightly integrated into the core clause in Persian than in German, occur most often in formal registers while in German they occur more frequently in informal registers (Lehmann 2024). Hence, we see that syntactic variants are recruited for register purposes based on their language-specific grammatical nature. Generally, mode and hierarchy appear to be stronger factors for register distinctions in our corpus design (see §3). Further potential strands of investigation that are promising with the current typological setup of the languages included in Lang*Reg are outlined in §5.

2. Background

Many corpora aspire to provide “naturally occurring language” (McEnery, Xiao, & Tono 2006: 4) in “authentic settings/ contexts [...] intended to be representative for a particular language, variety, or register (in the sense of reflecting all the possible parts of the intended language/variety/register)” (Gries & Newman 2014: 258), but “corpus compilation [can be] driven by a variety of considerations, some of which necessitate the inclusion of non-authentic texts and text types” (Barth & Schnell 2021: 8). A corpus can therefore more broadly be defined as “a collection of texts when considered as an object of language or literary study” (Kilgarriff & Grefenstette 2003: 334). In fact, corpora benefit from any kind of variety as long as different types of data included in a corpus are strictly separated or labelled so that a user

¹ We wish to thank in particular all the participants in the various regions as well as the collaborators on site who helped us with the complex procedure. We further wish to thank our transcribers and annotators, in no special order: Hadis Shirvani, Maryam Barani, Shahla Sadeghi, Sri Sumaryani, Maftuch Fahman Al Amiqi, Yolanda Marcela Binambuni, M. Dany Faiz, Joyner Hadinoto, Sina Andreas, Luka Anlauff, Florian Deichsler, Victor Renard, Henri Schellberg, Zacharias van Stek, Amedee Colli Colli, Edber Dzidz Yam, Motahareh Sameri, Mehrnoosh Taherkhani, Negar Ilghami, Mahdiye Arvin, Ameneh Ebadi, Sina Homay, and Prof. Mahmood Bijankhan.

can select what type of data best serves their purposes. The more diverse the available data of any given language is, the closer we get to achieving comprehensive documentations of languages (for a broad categorization of aspired data types in language documentation, see Himmelmann 1998; Lüpke 2010; Woodbury 2014). Endeavours for thorough labelling and annotation of situational-functional factors and metadata also help in making corpora “not a short-term record for a specific purpose or interest group, but a record for generations and user groups whose identity is still unknown and who may want to explore questions not yet raised” at the time of creation (Himmelmann 2006: 2). In this sense, any data made available, irrespective of size, can be of use to the community if properly documented, described and labelled.

For languages with a larger user base, the range in available data and contexts is continuously expanding. Take COCA (Davies 2007) and the BNC (Davies 2004) for English, which include language data from all kinds of contexts such as web, newspapers, novels, academia, conversations, etc., even though a strong focus is still put on the written medium (around 90% of the data) as well as publicly available data that tends to have very specific functions such as conveying information or entertaining. In other cases, the focus is primarily on spoken data types, though none of the existing corpora of spoken language – not even for better-studied languages – represent the range of variation existing in spoken language use. This is partly because of the necessity for elaborate efforts in transcription involved which up to now has limited the size of spoken language corpora. However, efforts in expanding spoken language resources have been increasing. For instance, the databank “Datenbank für gesprochenes Deutsch (DGD)” (IDS 2023), a collection of various spoken German corpora, has managed to compile an extensive number of conversations in various contexts, including public speeches and discussions, dinner conversations, game night interactions, meetings in various settings, conversations during café visits, etc. There have also been successful efforts to create multilingual corpora for cross-linguistic investigations such as the Language Documentation Reference Corpus (DoReCo: Seifart et al. 2022) and the Multilingual Corpus of Annotated Spoken Texts (Multi-CAST: Haig & Schnell 2022), both of which combine independently collected spoken data from a range of languages and ensure uniform annotation standards. The universal dependencies (UD) project similarly ensures cross-linguistic compatibility for treebank annotations through its universal yet still flexible annotation guidelines, and receives constant additions through open community efforts (Nivre et al. 2020). While these collections allow researchers to compare languages via their standardized annotation schemes, they are not comparable corpora in the sense of McEnery & Xiao (2007: 20), who state that “a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness [...] e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period.” Naturally, comparable corpora benefit from annotation standardizations because using the same annotation schemes strongly increases the practicability of these linguistic resources. A comparable multilingual corpus that incorporates uniform annotation standards across data sets is the RUEG corpus (Wiese et al. 2019a), for which speakers of various heritage languages reported an accident once to the police and once to a friend. This is one example of a corpus which also tackles intra-speaker variation as an operative factor similar to Lang*Reg across those two communicative situations. The SCOPIC corpus collection method (Barth & Evans 2017) also tackles comparable cross-linguistic data collection by using picture tasks in combination with description tasks and problem-solving tasks in four phases in a variety of languages. While neither RUEG nor SCOPIC are parallel corpora in which parallel texts in different languages would have exact semantic equivalence, via translation (Barth & Evans 2017: 1, McEnery & Xiao 2007: 20), they go one step further than comparable corpora which just sample similar texts from different languages by using the same task-based collection method across languages. This way, the resulting data is cross-linguistically much more uniform because the comparable task provokes speakers to use “broadly comparable formulations” (Barth & Evans 2017: 1) and talk about similar contents.

Lang*Reg complements such efforts by using a method that focuses on similar situational-functional settings across different languages in their respective cultural and geographic context, creating approximated natural environments for the conversations prompted therein (see §4) as well as targeting specific experimental variables, both situationally and with respect to interlocutors (see §3.2.1 and §3.2.2). Lang*Reg also aims to enhance documentation efforts, not only by adding new resources with a variety of contexts for the currently included languages, but also by showcasing a cross-culturally applicable data collection procedure that concentrates on intra-individual variation in a series of naturally occurring situations. Lang*Reg includes four different contexts and three types of communicative activities as well as two modes, representing different potential registers (Lüdeling & Kytö 2009; Pescuma et al. 2022) that can be compared across all five languages collected so far. Lang*Reg is unique in particular for the systematic inclusion of varying degrees of spontaneity in spoken varieties. The intra-individual design whereby the same language user is recorded in six diverse situational-functional contexts (see §3 for details) captures different levels of spontaneous lan-

guage output of that language user. With a minimum of 12 speakers for each language who were each recorded in all six contexts, it is possible to study similarities and differences in how individuals change their communicative behaviour based on the situational-functional context. The methodological design developed for Lang*Reg data collections allows us to include a variety of common situations that are cross-culturally relevant for language users, for example, unscripted, spontaneous conversations with friends and strangers. Next to the parallel data collection method, parallel transcription and annotation guidelines further enhance the compatibility of the data sets included in Lang*Reg. The corpus can therefore be used for many different purposes, allowing explorations across different kinds of disciplines. With this, Lang*Reg addresses some of the main documentation goals: diversity, (fairly) theory-neutral processing and standardization across data sets (see §4), and collaboration as well as accessibility with its archiving and public access approach (Himmelmann 2006; Holton 2012; cf. Woodbury 2014).

The exact Lang*Reg make-up and procedures will be detailed in the following sections. In §3, we present the corpus design with a specification of the six communication events, their situational features, and other characteristics of the data included in Lang*Reg. The contextual variables are detailed in §3.2, specifying external situational variables in §3.2.1 and interlocutor variables in §3.2.2. The procedure for the presented data collection method and the steps for processing the data are delineated in §4, supported by a walk-through of an exemplary routine in the Appendix. §5 describes the current language set included in Lang*Reg and how the constellation of cross-linguistic aspects holds numerous possibilities for comparative investigations. Lastly, we summarize the benefits of the methodology and the resulting corpus in §6.

3. Corpus Design

3.1 Text formatting and fonts

The Lang*Reg corpus is designed to represent the speech of the same speaker in different situational-functional contexts. For an overview of the specific characteristics see Table 1. A minimum of 12 participants per language² (for participant characteristics, see §3.2.2) traversed a course of six situations in which they were asked to produce language in two types of activities: a) telling a story to a friend, b) talking freely with various interlocutors (friend, stranger, taxi driver, university professor). Moreover, our design included the activity storytelling in two modes, which allows for the comparison between spoken and written language by the same language user. The data was collected in countries and locations where the language is usually spoken. Participants were recruited on site³ and recorded with lavalier microphones while engaging in all or most of the six situations on the same day. Recording sessions lasted 15 minutes each (except for the storytelling, which lasted only 2 minutes). An exemplary procedure looked like this: a participant starts telling a story to the friend before speaking to the same friend freely, followed by the participant speaking to a stranger, before entering a taxi and driving around while talking to the taxi driver for 15 minutes, and thereafter meeting the professor in their office, until finally writing down the story narrated previously to the friend (for details on the procedure see §4.2).

² For Persian, we added an additional set of eight participants of a younger age group (20-30) in order to test our assumption that age is an influencing factor with respect to register variation in this language. For Kurdish, we tested 20 participants in total because we included the parameter gender in the design, including a male and female professor as well as a male and female taxi driver as interlocutors, given that gender is a crucial factor in this language-cultural context. As can be seen, each additional factor requires an increase in the number of participants. This shows how the design can be adapted to specific needs of the socio-cultural context while ensuring comparability across data sets.

³ The researchers or collaborators living in the locations were tasked with finding participants before data collection started, using mailing lists, word of mouth and referrals by colleagues for recruitment. Attention was paid to the range of contacts so that not all participants were recruited from the same or similar sources.

Table 1: Characteristics of the Lang*Reg corpus design, detailing the six recording situations with their varying communication events (different interlocutors and monologue or dialogue type of interaction) and situational features (mode, length and space). Two main social relation criteria are varied: social distance between interlocutors (related to level of acquaintance) and social hierarchy between interlocutors, whether the participant (P) or their interlocutor (I) exerts more power in the scope of the activity.

No	Communication event		Situation features			Social relation	
	Interlocutor	Interaction	Mode	Length	Space	Distance	Hierarchy
1	friend	monologue	written	–	private	close	P = I
2	friend	monologue	spoken	2min	private	close	P = I
3	friend	dialogue	spoken	15min	private	close	P = I
4	stranger	dialogue	spoken	15min	private	distant	P = I
5	taxi driver	dialogue	spoken	15min	non-private	distant	P > I
6	professor	dialogue	spoken	15min	non-private	distant	P < I

By taking cross-linguistic as well as cross-cultural differences between speech communities into account, Lang*Reg allows for meaningful comparisons. To be able to achieve this, the recorded situations require cross-cultural validity, meaning they must occur naturally in a diverse set of cultures and languages. The Lang*Reg situations can be described with the following main parameters (similar to features found in the “context of situation” framework used in Systemic-Functional Linguistics, see e.g. Halliday & Hasan 1989; Neumann 2014): (i) hierarchy vs. equality, (ii) distance vs. closeness (implemented by us as acquaintance vs. non-acquaintance), and also (iii) mode, spoken vs. written, and (iv) monologue vs. dialogue. Each situation can thus be seen as a specific combination of these parameters.

In order to vary the two social relation parameters hierarchy and distance, very particular situations are required (see also §3.2). It may be pervasive in societies to have close and distant social roles with communication partners, acquainted as well as unacquainted interlocutors, but the range of situations where language users engage in conversations involving a hierarchical relationship varies between cultures, due to varying authority concepts. For example, in some cultures a village elder may have the most authority in contexts such as a village council meeting, whereas the spiritual leader receives more respect in a spiritual ceremony. In contrast, as the authority of the established churches has declined in postindustrial societies along with a shift away from authority attributed to external institutions, instead foregrounding the individual self, a spiritual leader might not incite a strong sense of hierarchy in such cultures (Inglehart & Welzel 2005), see also World Values Survey results (Haerpfer et al. 2022). Another factor constricting cross-culturally valid situations with hierarchy distinctions is that authoritative figures may entail negative connotations in certain locations, such as police officers in Mexico, making such officers ineligible as interlocutors with higher status in the Lang*Reg design. Inquiries into the attitudes of language users in all our recording locations (Germany, Iran, Mexico, Java) found that intellectuals such as doctors and professors had the most positive virtue while establishing a hierarchical order via the authority ascribed to them due to their extensive expertise (see §3.2.2). Ultimately, we selected university professors as collaborators in all recording locations. As an alternative to a professor, another person of high standing with authority is also appropriate for this situation, as long as that person is able to insinuate a sense of significance in the cultural and situational scope, thereby presumably increasing the hierarchical difference. Researchers aiming to use the Lang*Reg method would have to similarly inquire the situation in the language context and decide which contexts would best invoke such hierarchical distinctions, which need not be commercial or professional but could, for example, be of a religious kind etc.

While the professor is assumed to hold a higher status in our situation compared to the participants, an interlocutor as a service provider is expected to reverse the hierarchical configuration because the beneficiary, in this case our participant, is assumed to hold greater power than the interlocutor who needs to solicit their services (Halliday & Hasan 1989: 57). For the service interaction, we selected a communication situation where participants talk to a taxi driver during a taxi drive because people are more likely to engage in spontaneous casual conversations in such a functional environment, compared to alternative professional situations such as exchanging goods at a shop, where the focus of the con-

versation would be directed by the goods and the exchange. In many contexts of exchanges of goods and services such as shopping or restaurant visits, conversations tend to be schematic and therefore less comparable with the other oral contexts in the corpus.

As a result, the relationship between interlocutors is a vital concept in the corpus: participants were either acquainted with their communication partner – friends or otherwise in a close relationship such as partners or family members⁴ – or they were unacquainted and therefore complete strangers. This marks the parameter social role relationship. Our design included one conversation between friends, while the other three conversations were unacquainted interlocutors, including the taxi driver and the professor. This design produced recordings of spontaneous and natural conversations in varied cultural contexts.

In addition, we aimed at including the written mode for those languages whose users are used to writing – this holds for all languages which are currently part of Lang*Reg. In some of the Lang*Reg languages, the written medium is used extensively in the language community, such as in German or Persian. Other Lang*Reg languages are like Yucatec Maya: language users rarely write in this language, preferring the Spanish language for writing purposes (see §5) so that there are fewer contexts where writing in Yucatec Maya feels natural. One context in which writing appears to be widespread even in smaller language communities is in private communications such as letters and nowadays in particular instant messages or online communication. This is the case for Javanese, which is similar to Yucatec Maya in that it is less commonly used in the written medium in formal situations, but is actively used in written private communications, especially online. Motivated by being a cross-culturally valid written context and a communicative situation that most language users will encounter in their daily lives, participants were asked to write a personal message to an acquainted person. In order for both modes, spoken and written, to be highly comparable, participants received the same task, to tell a story, and produced the same content in both the spoken and written communication event; meaning they told the exact same story one time to their friend orally and one time as a written message in the form of an informal letter (to the same friend) as if the friend hadn't heard the story before. The form of a letter guarantees compatibility with the oral storytelling event in terms of interaction, whereas a story told via an instant message would usually be more interactive. These two situations are represented in Table 1 in situation 1 (written storytelling) and situation 2 (oral storytelling).

Beyond the two storytelling events, Lang*Reg includes four conversations in which both interlocutors actively engage with one another. The first of these is a casual conversation with a friend, situation 3 in Table 1. This is contrasted by a conversation with a stranger that stands in no particular relation to the participant, situation 4. Furthermore, each participant talks to a stranger with whom they engage in a professional activity, thus having a hierarchical relationship in the event where the participant assumes a higher status: they experience a taxi drive as an exchange of a service in which they talk to the taxi driver, situation 5. Lastly, participants were instructed to talk to a professor in the professor's office, situation 6, where the professor asks the participants a series of questions about travelling during the pandemic, thereby assuming a higher status in the social hierarchy in contrast to situation 5. Unlike the other conversations, this type of conversation can at times resemble an interview due to the fact that casual conversations are not very natural between a professor and another person in a professor's office. Participants tended to seek the guidance of the professor in this context, to varying degrees, falling back on answering questions rather than talking freely or even asking questions themselves.

As a result, Lang*Reg includes six communication events in which the same participants were recorded. The situations in which the data collections took place were also tightly controlled for a series of further contextual and verbal variables to both heighten the authenticity of the communication events and also reduce effects of other potentially influencing variables – these are described in §3.2 and §3.3.

⁴ Throughout this article, when referring to the acquainted interlocutor, “friend” is used even though the relationship may be more intimate for some interlocutor pairs.

3.2 Contextual variables

In register studies, researchers have hypothesized a variety of situational-functional parameters that may influence the speech of language users and hence determine a register. In the Systemic-Functional Linguistic approach, these parameters are conceived of as part of networks, usually clustered into the mode of discourse, the material side of the interaction (“how”), the tenor of discourse, the relationship between discourse participants (and therefore “who”), and lastly the field of discourse, the content side (or the “what”) of the communication event (Halliday 1978; Halliday & Hasan 1989; Hasan 2014; Neumann 2014; Teich 2003). Others have also proposed situational frameworks for the classification of texts, for example Biber (1994), in which the situational-functional characteristics are given as grouped parameters to allow for the characterisation of situations at different levels of generality. All these approaches include parameters that describe the characteristics of participants, the relations between them, setting, channel, relation of interlocutors to the text, purposes and goal, and lastly the topic of the interaction.

These different approaches result in very similar sets of parameters. We strove to control for as many parameters as possible and vary the ones necessary to create authentic situations. First, we will describe the parameters concerning the setting and how the event unfolded, the time, place, channel, interactional features etc., in §3.2.1, followed by a specification of the interlocutor characteristics in §3.2.2. Lastly, we will delineate the content variables, what is talked about, in §3.3.

3.2.1 External situational variables

One of the central parameters in Lang*Reg as presented in Table 1 is the interactivity of the situation, which is firstly influenced by the presence or non-presence of more than one interlocutor but also by the type of activity performed (see §3.3 about activity). The first situation (written storytelling) is the only recording situation where the addressee is not present because language is mediated via the written mode, using a graphic channel of communication. Hence, there is no interaction between participant and addressee in the written monologue. There is also no interaction in the second situation (oral storytelling), although another interlocutor is present in the oral communication. The lack of interaction here stems from the goal of the activity – telling a story, where one participant tells a story while the other listens, so that phonic interactions between communication partners were minimal, with few acknowledgements or comments by the addressee. The oral storytelling situation is therefore also monologic. In contrast to the first two recording situations, the remaining situations 3-6 are interactive, as they constitute dialogues between two speakers.

Another external situational parameter varied in Lang*Reg is the mode of conversation, the physical channel used to convey language. In all situations except the written event (situation 1), communication takes place orally, using the phonic channel, in a face-to-face constellation.

The communicative space in the external world, the location, can also affect how participants behave (non-)linguistically. Four of the selected situations, namely situations 1-4 (written and oral storytelling, conversation with friend and with a stranger), represent communication events in the private sphere where language users meet family and friends and on occasion also people they have not met before. Therefore, the place of recording for these situations was deliberately chosen to emulate a private space such as a living room, a private room in a café, or a garden.⁵ Participants were able to sit down comfortably and were offered drinks and a small selection of snacks. Plants and decorations also helped to create a cosy and welcoming atmosphere. The space generally aimed at heightening the sense of privacy. Keeping the location for situations 1-4 the same across recording events further increases the relevance of the participant’s social relations.

⁵ Under COVID-19 restrictions, recordings in enclosed spaces were not always feasible so that alternative spaces where private interactions take place such as in a garden, preferably segregated, without the eventuality of passers-bys or over-hearers of any kind, cf. in particular Bell’s (1984) audience design, were accepted as well.

The remaining two situations are, in contrast, non-private interactions. Situation 5 (conversation with a taxi driver) focuses on the functional interaction resulting from the professional service provided by a taxi driver.⁶ As such a communicative situation occurs naturally during taxi drives, the recordings were also made during a taxi drive. This situation thus unfolds on public streets, yet due to the enclosed car interior, there are no over-hearers (see audience design by Bell 1984) similar to situations 1-4. Situation 6 (conversation with a professor) took place in an office(-like) location. In most recordings, the professor either had an office for the recording or we were able to lend an office, informing participants that the room they were about to enter was indeed the professor's place of work. Yet, in some cases, the pandemic did not allow us to record in an enclosed office space so that we opted for a location with a sense of formality. For Yucatec Maya in Mexico, we therefore recorded the professor interview in the garden of a museum and cultural institution. For Javanese in Indonesia, we recorded the professor interview at the professor's house, in the guest living room where it is culturally appropriate to have scholars visit for a professional meeting. The professor was further instructed to wear formal clothing, such as a suit or a culturally significant costume. The interaction in situation 6 was therefore supported by an official location and the professional demeanour of the professor which heightened the hierarchical distance and the sense of formality.

3.2.2 Interlocutor variables

As Lang*Reg aims at contrasting the different relationships between interlocutors based on a small selection of parameters, it was vital to reduce the influence of other social variables. Participants were therefore selected by a strict regimen of criteria.

3.2.2.1 Speaker Sample

The participants were between 30 and 50 years old. Restricting the age range minimizes effects from generational variation while the longer life experience increases the chance to learn a more refined understanding of register differences due to more opportunities for being exposed to varying situations. We found this age range to be an advantage to handle conversations with strangers (compared to a younger age range). For each language, the participant set consists of about half female and half male speakers (self-identified). Such a balanced set enables researchers to check for potential gender-related variations.

Concerning the level of education, the minimum requirement was set to 12 years of education, which corresponds to Level 3, Upper secondary education of the International Standard Classification of Education of 2011 (OECD & Development 2015). Variation in the social background is best kept minimal to guarantee largest comparability between speakers, reducing social distance arising from factors such as education. The level of education is used as an indicator for the social categorization of participants. Setting the level of education to a minimum of 12 years means that participants are more likely to have been exposed to a similar range of registers on the social dimensions of distance and hierarchy – setting the boundaries of familiarity and formality (Ravid & Tolchinsky 2002) – due to their education. It also ensures that they belong to a similar subset of the general population, thereby minimizing social variability in the data.

There was no requirement on the occupation (status); however, such information was recorded via a sociolinguistic survey following the International Standard Classification of Occupations (ISCO, International Labour Office 2012), which classifies ten major groups. All speakers also should have lived in the same dialectal area for a large part of their lives; this is important to reduce effects of dialectal variation.

⁶ The taxi driver and professor were contracted prior to data collection to act as interlocutors for the participants in those situations. The taxi was provided by the taxi driver. For more details on the procedure, including consent management, see §4.

3.2.2.2 Variables of interest

The parameter that varies by design is the social relation between interlocutors. The main reasoning behind this variable is that the depth of personal common ground between speakers increases with the level of acquaintance (assuming a four-degree differentiation: strangers, acquaintances, friends, intimates). While strangers have no common ground prior to their first interaction, acquaintances will already have established a shared basis from their shared experience. This shared experience is even larger for friends and presumably largest for intimates (Clark 1996; see also Krifka 2008). Here we implemented a two-level distinction between no common ground (strangers or unacquainted participants) and greater common ground (friends and intimates). As participants in the friends and stranger situations are in a reciprocal relationship, both being friends or both being strangers to each other, there is no hierarchical distance in their social relation.

Supplementary to the participants with the characteristics as defined above, it was necessary to sign on two collaborators per language as not-at-issue interlocutors to support situations 5 and 6. These conformed as closely as possible to pre-defined characteristics, both collaborators were to be unacquainted with the participants. According to the ISCO, which indicate prestige and socioeconomic status scales based on education and income (Ganzeboom & Treiman 1996), the professor collaborator (ISCO Code 2310) has a higher occupational status than the taxi driver collaborator (ISCO code 8322). As both appear in a professional role, their occupational status is indicative of hierarchical relations as well. The taxi driver provides a direct service, transportation. The assignment of roles in this business interaction grants the beneficiary of the service more power than the provider, resulting in a hierarchical distance between participant and taxi driver.

The professor also appears in a professional capacity in the recording event, even though they provide a more general, public service, education and cultural contributions. What distinguishes the taxi driver and the professor is the inherent sense of authority ascribed to culturally significant personae such as professors, which are recognized in most societies as exceptional in part because of their expertise and overall prestige (see about social power and ranking Brown & Levinson 1987: 74f.), resulting in authority of epistemic nature concerned with knowledge (Stevanovic & Peräkylä 2012: 298). The relationship between participants and the professor is thus characterized by an imbalance in authority, granting the professor more power because the perceived expertise of the professor will be larger than that of the participants, not just in their field of profession, but perceptively also in more general terms due to public conception (see cross-cultural dimensions in the World Value Survey, cf. Haerpfer et al. 2022). An effect of authority, defined as a function of socially-legitimated inherently unequal role relationship in Poynton (1985), may, for example, be observed in the linguistic output along the dimensions of mood or modality. A speaker might use more expressions of modality when talking to a person with a higher level of authority or there might be a higher frequency of demands and imperatives when the addressee has a lower level of authority (cf. Poynton 1985: 79, Neumann 2014: 137).

3.3. Verbal communication variables

In all the situations of the Lang*Reg design, language has a constitutive role of the situation, meaning that participants' primary reason for being in the situation is to communicate. Of course, the fact that speakers are partaking in a data collection makes this a staged communicative event (based on e.g. Himmelmann 1998; Lüpke 2010; Woodbury 2014), but one with a very low amount of control imposed by the researcher, meaning the way the researcher “shap[es] and manipul[at]es its structure and content” (Hellwig 2019: 17), because participants were only given a topic to talk about and were otherwise encouraged to talk freely. The lower the amount of control, the higher the naturalness of the situation, “the degree to which the event would have taken place even without researchers asking for it” (Hellwig 2019: 17). Despite the fact that these are events that occurred precisely because of the data collection, the recording situation was staged in a way that participants were more likely to forget the recording circumstances (see §4), thus increasing the chances that speakers act as if the event had occurred naturally and foregrounding the interaction with the respective communication partners.

The situations differ in the kind of communicative activity performed by participants. They told a story in situations 1 and 2, which is a specific form of narration, involving mostly past events. The storytelling activity allows participants to act freely with uninterrupted time to sort through thoughts. All other situations (3-6) required interlocutors to negotiate the communication space for speaking time.

Register is also defined by the topic of the linguistic exchange (Biber 1994; Neumann 2014; Teich 2003). In order to reduce (uncontrolled) variation between situations, language users should talk about the same things in each situation – yet in practice authentic interactions cannot be controlled so tightly, for natural conversations tend to take unforeseen turns and develop throughout the interaction. However, by setting a general theme for all the different interactions, it is possible to induce a similar frame of lexical material. Such a theme should ideally give grounds for natural conversations in each of the situational and cultural settings.

For our design, participants were encouraged to talk about the general theme of travelling or visiting places in all situations, from recounting a travelling experience in situations 1 and 2 to generally talking freely about travelling experiences, plans, and wishes in situations 3 and 4. This topic also came natural to the taxi drive interaction, for people usually use taxis when travelling to and from places, be it short-distance or as part of long-distance travels. The taxi driver was also able to use the surroundings during the journey to incite an interaction. The communication with the professor, on the other hand, focused more specifically on the very timely topic of travelling or moving around under COVID-19 restrictions and consequences for participant's ideals or goals from the experiences during the pandemic. This made it a plausible interaction with a professor who was perceived as specifically interested in this topic.

4. Procedures

We conducted the Lang*Reg data collection in places where the language is naturally spoken: Berlin, Germany for German; Tehran, Iran for Persian; Ilam, a town in the west of Iran, for Ilami Kurdish; Felipe Carrillo Puerto, a city in Quintana Roo, Mexico, for Yucatec Maya; and Semarang, a city in Central Java, Indonesia for Javanese. The following sections describe in detail how the data was collected and what measures were taken to ensure both a natural space for language use as well as a controlled environment for the variables at issue.

4.1 Informed consent

All participants received a subject information sheet to keep prior to data collection, with general information about the project and the researchers' affiliation as well as contact details, also including what data is recorded, how it is recorded, and how the data will be used. They were further informed that their participation is voluntary, that their data will be treated confidentially, and that they are to be compensated⁷ for the parts fulfilled (for every 10 minutes completed). Participants could stop the procedure at any moment without repercussions. Participants are also able to revoke their consent at any moment, at which point the data will be deleted irrevocably. They created a unique code that is easy to recreate for an individual as it is based on criteria such as family members' first names using predefined letters (first and third letter of the mother's name) and their or their family member's day of birth (day only) to ensure that they can request deletion of the data even post-anonymisation. Personal information will be deleted at the end of the project or at the latest ten years after participation. They further received contact details to the data protection officer of the project's home university.

Following a discussion of the subject information sheet, giving ample opportunity to ask questions, participants were asked to fill out a detailed consent form in which they specify their agreement to be recorded via microphones, provide handwritten language and personal information as part of a social questionnaire, agree that their non-biometric data will be stored in research repositories, that the anonymised results from the data collection will be made available to the public outside of the project, and that the recorded data can be used for scientific purposes by researchers of the project as well as researchers outside of the project. The collaborators contracted as interlocutors, the professor and taxi driver, received the same information and consented to the recordings prior to taking part in the data collection.

Researchers took particular care to make sure all participants understood that their participation was voluntary and that their data is treated most confidential. This was important especially in the case of some groups that had never participated in a research project before, or were unfamiliar with the informed consent process.

⁷ Standard rates based on the location.

4.2 Main data collection

For one recording day, we invited four participants, two sets of friends who were not acquainted with the respective other pair. This way, we were able to record situations 1-3 (communication events between friends) twice on one day as well as situation 4 twice (communication between strangers) by pairing each participant of friend group A with a participant of friend group B. Making sure that participants of group A were unacquainted with those of group B was a particular challenge in smaller communities where language users are mostly familiar with each other. In some rare cases, the unacquainted interlocutors had seen each other before but had not engaged in more extensive interactions. Situations 1-4 (written and oral storytelling, conversation with friend and with a stranger) were thus all recorded on the same day for each participant. Situation 5 (conversation with taxi driver) and situation 6 (conversation with professor) were either also recorded on that same day or on an adjacent day if scheduling required it. For each language, we had at least three recording days. These were all recorded in the same city for each language.

Between the oral storytelling situation and the written storytelling event was at least one other recording situation in order to ensure that participants were not overly primed by their previous retelling. Situation 1 (written storytelling) was usually the last task in the data collection, unless participants had a longer waiting period in which they could do the assignment. The order of the dialogue situations (3-6) was reversed for half of the participants so that one half started with the audio recordings in the private setting before moving to the non-private locations while the other half started recording in the non-private situations before settling into the private location. This means one group, for example, started with the conversation between friends and proceeded step-by-step to the conversation with a stranger, then the taxi driver and lastly the professor while the other half started with situation 6 (conversation with professor) and traversed the course of situations towards situation 3 (conversation with friend). For scheduling reasons, it was sometimes more efficient to switch the recording order within the private sessions, the conversation with a friend before or after the conversation with a stranger, or within the non-private sessions, the conversation with a taxi driver before or after the conversation with a professor. However, participants always recorded all the oral communication situations in the private location together, and recorded all the situations in the non-private locations together.

Before each recording event, participants were instructed that they were about to enter a situation in which they should talk freely with their interlocutor for about 15 minutes. Although they were given a topic for the conversations, participants were explicitly told to talk about whatever they wished, meaning the theme of travelling as described in §3.3 was only a starting point and general frame. Due to the fact that many language users in our cultural settings use more than one language on a daily basis, participants were specifically not told which language to use. If the participant asked whether they should speak in a particular language, the researcher told them that they should use whatever felt normal in the situation. This being said, participants were recruited with the requirement that they speak the target language as a native language,⁸ so we assume that they came with the expectation that they would use the target language.

For the recordings, lavalier microphones were attached to the participant and recordings started before they entered the situation. Participants had no contact to their unacquainted conversation partners prior to the recording event. For situation 2 (oral storytelling to the friend), participants were instructed to tell a two-minute story about a travelling event (see §3.3) that the friend had not heard before. After two minutes, they received a signal to switch, a knock on the door or a quick approach by the researcher for outside recordings, after which the other friend began telling a two-minute story. Following the second round of storytelling, they received another signal, after which they engaged in the dialogue on travelling for 15 minutes. The researcher extracted themselves from the situations as soon as the recording started so that participants did not have the feeling of being observed (see the Observer's Paradox in Labov 2006: 100). This was achieved by either closing the door when a recording took place in a room or moving out of the field of vision for outside recordings. The same applied to the recordings between strangers, for which participant of group A was seated in the private location before the stranger from group B entered or approached the recording location while the researcher started the recording from outside or afar. Both participants were instructed to talk freely

⁸ In fact, we restricted the range of participants to those of certain dialects in each region, either controlled through the recruitment location and the amount of time lived in the region or by directly inquiring about speakers of a dialect. This restriction was to avoid possible effects of dialectal variation of linguistic phenomena.

about travelling with the unacquainted communication partner for 15 minutes, until the researcher gave a signal. To increase the chance that both participants engage in the conversation equally, they may both additionally be instructed to find out about the other's experiences or preferences when it comes to travelling.

Prior to the recording days, we signed on two collaborators who then functioned as the professor and the taxi driver for situations 5 and 6 respectively. They were tasked with initiating the dialogue, keeping the participants engaged in the conversation and steering the conversation to the general topic pervading through all conversations, which was moving around, travelling and places (see §3.3). However, they were not to enforce the topic strictly if the conversation developed naturally into another direction – for instance, the taxi driver and one participant in the German data collection ended up talking about sports for part of the conversation. The taxi driver picked up each participant at a pre-determined location, usually on the same day as recordings of situations 1-4, and took the participants for a 15 minute (minimum) drive through the area, engaging them in a conversation about travelling, places visited, and activities. Throughout the drive, participants mostly sat in the back seat. The researcher was not present in the car. After a participant had entered the car, the researcher switched on the lavalier microphones and started the recording. For the dialogue with the professor, participants entered the office-like location as described in §3.2.1 with the lavalier microphones turned on. The professor then started asking them questions about travelling during the pandemic and consequences resulting from this for the participants.

For the written assignment, participants were given pen and paper and were instructed to write down the story they had told previously to their friend as if they were writing a letter to this friend, telling the story for the first time.

An exemplary walk-through of the whole procedure including timings and instructions is provided in the appendix in Table 2.

4.3 Surveys

In addition to the six recording situations, participants also filled out an extensive social questionnaire on a computer or iPad⁹ using LimeSurvey, which they either completed by themselves or with the help of a researcher. LimeSurvey allowed us to ask a series of conditional questions based on the languages spoken by each participant, which considerably reduced the size and complexity of the survey for participants, since questions or parts of questions can be hidden if not relevant. In the questionnaire, participants provided information about their personal, educational and vocational background, socio-economic details, their social network at home and when growing up and their history with languages, including contact situations, rate of use in context with different people and general language skills for various media types. This task took about 45 minutes. Group A was given the questionnaire when group B traversed the non-private sessions with the taxi driver and the professor, and group B worked on this task when group A was recording the non-private session. Alternatively, the questionnaire can be completed before or after the main data collection.

At the end of the data collection day, all participants filled out a very short post-study survey on paper, with the questions in (1). They were asked what they noticed during the recordings (1a) and what they thought is being researched (1b). Also, the survey inquires which parts they found more difficult or easier (1c) and whether they employed any strategies (1d). Lastly, it inquires as to how their impression of the other participants they had talked to was (1e).

- (1) a. Did you notice anything during the study, such as patterns, regularities, something surprising or strange?
- b. What did the study investigate according to your opinion?
- c. Which parts of the study did you find particularly difficult or easy?
- d. Which strategies did you employ – if any – during the study?
- e. What were your impressions of the other participants you talked to?

In general, the answers by participants indicate that they did not notice anything unusual during the data collection. On the contrary, most participants seemed to enjoy the parcours of conversations, with some stating that they found

⁹ As data collections took place in the field and internet connections are not always stable, a paper version of the survey was available, though the paper version was highly complex and long due to the many back references, for example when asking for details about how, when or with whom languages previously stated are used.

it very interesting to observe how they use their language with different people. From the survey, it became clear that some participants noticed that our interest of research was language spoken in different contexts, yet we do not believe that this affected how they engaged in the recording situations. Participants from languages such as Yucatec Maya and Javanese were very happy for the interest shown in their language, saying that they felt proud. The data collection was generally perceived as very professional. It was easy for most participants to record the oral conversations whereas many found the writing task challenging, even in cultural contexts where writing is pervasive in the native language. The second difficult part, according to the post-experiment survey, was the social questionnaire, which participants perceived as long – this effect appears to be reduced when a researcher guides them through the questions instead of participants doing it by themselves. We also noticed some cultural differences; for instance, Javanese speakers commented that they found it harder to start conversations with unacquainted interlocutors. Persian speakers, on the other hand, found the conversation with the professor the most difficult of all the oral tasks as they felt less comfortable due to the prestige of the professor. The same was not commented about (as much) in other cultural contexts.

4.4 Data processing

All the recorded data has been transcribed using ELAN (ELAN 2022) by trained native speakers of the respective languages. The audio signal received a close transcription for each respective time span. A close transcription is a one-to-one transcription where the audio signal is transcribed directly without making changes to what can be heard, without adding elements, deleting elements etc. (Wiese et al. 2019b). This way, the transcription reflects repeated content and incomplete words, phrases or sentences. Transcribers used a basic syntactic segmentation, meaning that one matrix clause and all its dependent clauses are transcribed in a single segment. A segment refers to the time-aligned span that determines where a speech unit begins and ends. Utterances smaller than a matrix clause also constitute one speech unit and are therefore segmented individually: one segment per independent phrase, one segment per interjection, one for a response particle.

Due to the different types of data (written and spoken) with different conventions and signals, a normalization layer is added following the transcription in order to facilitate search and automatic annotation processes. On this level, the orthography is adapted to the standard spelling or conventionalized practice. This refers specifically to lexical spelling as the normalization is applied mainly to the word-level (Wiese et al. 2019b). For each language, a dictionary, lexicon or similar work of standardization has been selected for this purpose, such as the current Duden version for German (*Duden – Die deutsche Rechtschreibung* 2020), or for Javanese, where there is no official standard spelling, the spelling conventions in the Javanese-English dictionary (Robson & Wibisono 2002). The normalization layer helps automatic taggers to correctly analyze tokens and enables researchers to find all instances of a word irrespective of spelling inconsistencies or mistakes, in the written mode, or very close variants in the spoken mode because in the normalization, they are represented equally. The advantage can be exemplified with contractions as in example (2a) from German where the second person singular pronoun is cliticized onto the verb while in other instances the pronoun might be independent, see (2b). The transcription layer will represent the difference between contracted and independent form, but for further analyses, the normalization layer will represent both cases equally, as in example (2b). Similarly, different spelling conventions are streamlined via the normalization, as for instance in Javanese, where word-final glottal stop is usually written as <k>, as in *mbak* ‘Miss’, yet in instant messages or informal writing the glottal stop is not necessarily spelled, as in *mba*.

- (2) German
- a. contracted form
- hast=e*
- have:2SG=2SG
- ‘you have’
- b. independent form
- hast* *du*
- have:2SG 2SG
- ‘you have’

These main steps were conducted for all the sub-corpora in Lang*Reg. Further language-specific annotations were then added, glossing for Persian, Yucatec Maya, Kurdish and Javanese, while lemma and POS-tags were used for German, which can be tagged automatically using the UDpipe parser (Straka, Hajic, & Straková 2016). A syntactic annotation with respect to clause types and constituent type based on Universal Dependency (UD: Marneffe et al. 2021) has been applied for Persian and German.

The data has been converted from ELAN to the browser-based search and visualization architecture ANNIS (Krause & Zeldes 2016). ANNIS allows for using complex queries as well as visualising the data of each data set in the corpus individually or all the data sets in Lang*Reg combined. Queries may include the extensive metadata provided about the situation and the participants.

5. Language Sample

Lang*Reg is specifically designed to investigate the impact of cross-linguistically varying factors on register variation including both linguistic and extra-linguistic aspects implemented through the choice of the sample languages. The current sample includes languages which systematically differ regarding a number of typological properties instantiating at the same time widely different socio-cultural situations and research histories. Although the current choice of languages is non-exclusive, meaning that further languages may be added in the future, the Lang*Reg corpus already covers various important aspects in view of our research aim to develop a cross-linguistically valid model of register. For one, it covers both largely monolingual, German speakers from Berlin and Persian speakers from Tehran, as well as multilingual speech communities, Yucatec Mayan speakers from the Mexican peninsula Yucatán, also competent in Spanish; Javanese speakers from the Indonesian island of Java, also competent in Indonesian; Southern Kurdish speakers from the mainly Kurdish-speaking Iranian province Ilam, also competent in Persian. Lang*Reg further includes dominant languages (German in Germany, Persian in Iran) and also languages with a minority status (Kurdish in Iran, Javanese in Indonesia, Yucatec Maya in Mexico), wherein Kurdish and Persian are in a unique position as both the minority and dominant language are represented in the Lang*Reg corpus. Furthermore, Lang*Reg covers not only minority languages closely related to the dominant language – see Southern Kurdish and Persian which are closely related Iranian languages as well as Javanese and Indonesian which both belong to Western Indonesian, Malayo-Polynesian languages – but also Yucatec Maya which is, as a Mayan language, genetically unrelated to Spanish. With such a diverse set of contact situations, Lang*Reg allows the comparative study of the role of register in variation and ongoing contact-induced change (cf. D’Arcy & Tagliamonte 2015). We can, for example, investigate how Spanish obligatory (inflectional) plural marking influences the use of the optional plural marker *-o’ob* in Yucatec Maya discourse. Lang*Reg makes it possible to observe overt plural marking in different contexts, potentially reflecting processes of change with respect to the optionality of plural marking. Based on previous register studies on language change we expect that rates of overt plural marking are higher in more formal registers, in situations with more distant and hierarchical social relations between interlocutors.

Next to the differing language contact situations, the current language sample instantiates crucial social-cultural differences with respect to language use in distinct contexts. This is illustrated with the variation of how spoken vs. written modes are used in our language sample. Persian has a high degree of literacy and a rich written tradition with elaborate norms for written vs. spoken varieties, and high usage in administration, education, and media. German is similar regarding the richness and diversification of the language, being used in all areas of communication, both in written and spoken contexts. In the case of Javanese, the contact language Indonesian is primarily used in all formal spheres, education, government, media, etc., while Javanese is used for informal oral or written communication, with some marginal representation in media (see Sneddon 2003a). Yucatec Maya is primarily used in oral communication in rural communities with low literacy in the population (Pfeiler & Zámešková 2006); yet Yucatec Maya has recently been introduced to the education system and has an evolving literal use that has been accompanied by processes of establishing norms promoted by academies. Southern Kurdish as spoken in Ilam is mostly used in colloquial speech with limited usage in formal situations and little recent use in literacy despite a literate history; instead, the dominant language Persian is used for formal contexts, education, and media (except for some limited local channels), and Persian is also the standard in all written contexts. Among the minority languages currently in Lang*Reg two thus have a limited writing tradition, Yucatec Maya and Southern Kurdish, while one has a broader writing tradition – namely Javanese. Present-day Javanese uses the Latin writing script and is primarily written in non-formal spheres, often with its own shorthand code in instant messages similar to Indonesian sms and Twitter (Brugman & Connors 2019). Speakers of Ilami Kurdish are usually unfamiliar with both standard Latin-based and Arabic-based alphabets (commonly used for other varieties of Kurdish such as Sorani), so they transliterate their sentences into Persian script in their limited usage of Kurdish for instant messages or colloquial writings. The different writing traditions might result in stronger distinctions between the written and spoken situations in Lang*Reg for Persian and German whereas the written situation in particular in Yucatec Maya and Kurdish could resemble more the spoken situations – on the other hand, it is also possible that speakers of such minority languages transfer their register knowledge from the dominant language. This will be one avenue of research stemming from the current Lang*Reg corpus.

Lang*Reg also lends itself to investigations into the diversity of registers and how register diversity may be influenced cross-linguistically by normative aspects. Because the Lang*Reg methodological design promotes spontaneous conversations for several of its situations, it allows for speakers to use the linguistic devices or the language that is relevant to those situational-functional parameters. Our language sample is illustrative of a number of these factors that directly affect register. Although both Persian and German show similar properties regarding language use in education, administration, media etc., Persian has been shown to have more salient differences between registers, diglossia (Ferguson 1959; Modaresi-Tehrani 1978). In contrast, Javanese is well-known for its three speech levels, *krama* ‘High Javanese’, *madya* ‘Mid Javanese’ and *ngoko* ‘Low Javanese’, wherein each level is represented by an independent lexicon and has specific morphology and morpho-phonological processes (Poedjosoedarmo 1968). At the same time, there are other factors at play given the multilingual contact situation, present-day Javanese speakers are almost all bilingual with Indonesian (Sneddon 2003b). Given the role of register in each language, Javanese and Indonesian are arguably in a diglossic state, where Indonesian is used in all formal situations. Indonesian itself represents diglossia, with ‘Standard Indonesian’ used in newspapers, formal public speeches, etc., while a local variety of Indonesian is used elsewhere (Sneddon 2003a). Indonesian does not have the extensive speech level system like Javanese, but does encode a pronominal distinction with first and second person, *Anda* (formal) vs. *kamu* (informal) for second person. Within Javanese discourse, speakers often code-switch to Indonesian, especially when certain functional-situational parameters would call for the use of *krama*, high Javanese, such as when speaking to a stranger (Goebel 2005; Nurani 2015). Yucatec Maya, on the other hand, is a language with less pronounced register contrasts; in this, it is comparable to Kurdish, which as another minority language with more limited contextual usage does not have such a clear distinction in different registers as Persian with its explicit lexical and grammatical variants that mark different levels of formality. In sum, we expect languages such as Persian, Javanese and German to show more fine-grained register differences between all of the situations in Lang*Reg while Yucatec Maya or Kurdish might generally be more similar across the situations.

Lang*Reg lends itself particularly to the study of high-frequency grammatical alternations which can be examined for their register sensitivity across the sample languages, such as alternations in word order, voice, or case (the dative alternation, the genitive alternation; see, among others, Grafmiller 2014; Szmrecsanyi 2019; Szmrecsanyi & Engel 2023).

Another topic highly relevant for register distinctions relates to the variable expression of grammatical categories such as number, case, or TAM, which are optional (to some extent) in several languages. For instance, in Persian, German, and Javanese, future marking in order to signal future time reference is optional: these languages have explicit gram-

matical future markers and underspecified forms (present tense markers in Persian and German, or bare predicates in Javanese). Based on this variation, the role of register in the distribution of explicit marking vs. underspecification can be investigated across these languages. The working hypothesis to be examined from a cross-linguistic perspective is inspired by Persian: In Persian, overt future marking is proposed to categorically indicate a high or formal register, while the underspecified form is used elsewhere (Ghafar Samar & Bhatia 2017). More broadly, this research strand tackles the question: How cross-linguistically valid is the pattern that more formal registers use more explicit expressions of a linguistic feature (see Biber & Finegan 1994: 317)? This hypothesis predicts overall higher rates of explicit future time reference across languages in the interview situation with the professor, while underspecified forms are expected to be most frequent in the situations with friends (1–3) because they feature the most shared context. At the same time, what are the implications of (potential) meaning differences for the register-associated use of explicit vs. underspecified forms?

In a similar vein, our sample languages allow for examining the impact of register on pronoun realization. The sample languages instantiate different types of null arguments: Yucatec Maya is a language with obligatory cross-reference marking for both subjects and objects with independent pronouns not normally occurring in the core clause; Javanese and Persian have been characterized as radical prodrop (or discourse prodrop) languages (Sato 2015; Sato & Karimi 2016), whereas German allows null arguments only as topic drop (Cardinaletti 1990). These typological differences are related to systematic variation in overt pronoun rates across these languages. Furthermore, given that pronoun realization is immediately related to discourse context, which varies by situational context (Schnell & Barth 2018), it is to be expected that registers will differ in their pronoun rates – interacting with the null argument type of a language. The relevance of register-related parameters for argument drop has been discussed for German (Schäfer 2021) and Persian (Haig & Adibifar 2019). These are just a few examples to illustrate the type of studies that can be fruitfully undertaken with the Lang*Reg corpus.

6. Conclusion

We have shown how Lang*Reg and its corpus design implemented through a comprehensive methodology enable researchers to take intra-individual variation, diverse contexts as well as cross-linguistic aspects into account. Lang*Reg further extends the available corpus and documentation landscape by looking at the same language users across multiple contexts and a uniform methodology across languages. We have presented a series of pointers for potential comparative research strands that could be fruitful with Lang*Reg. Lang*Reg can, of course, be a great resource for numerous further angles of research. The complete corpus is available under a public domain license (CC-BY-NC-ND) in Zenodo at <https://doi.org/10.5281/zenodo.7646320>.

Ethics and consent

All data collections adhered closely to the ethics standards in Germany and were approved prior to realisation by an ethic committee. Participants were given ample information about their data privacy rights. All participant consented to the data collection and the use of their data as described in this article. See §4.1 for more details.

Funding information

The corpus has been developed in the course of the research undertaken within the CRC 1412 “Register: Language Users’ Knowledge of Situational-Functional Variation” at Humboldt-Universität zu Berlin and the Universität zu Köln. The research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC 1412, 416591334.

Competing interests

All authors have read and agreed to the published version of the manuscript and have no competing interests to declare.

Authors’ contributions

Conceptualization: A. A., E. V., N. L.; Data collection protocol / methodology: N. L., V. M., D. M.; Data collection & curation: N. L., V. M., Z. F., J. V. K., D. M.; Writing manuscript (original draft): N. L.; Writing language-specific descriptions (in particular §5): E. V., J. V. K., V. M., Z. F., N. L.; Writing revisions/review editing: N. L., J. V. K., E. V.; Supervision, project administration, funding acquisition: A. A., E. V.

References

- Adli, Aria, Elizabeth Verhoeven, Nico Lehmann, Vahid Mortezapour, & Jozina Vander Klok. (eds.). 2024. Lang*Reg: A multi-lingual corpus of intra-speaker variation across situations. Version 0.2.0. [Data set]. Zenodo. Berlin: Humboldt-Universität zu Berlin. doi.org/10.5281/zenodo.13889198
- Barth, Danielle & Nicholas Evans. 2017. SCOPIC Design and Overview. In *Language Documentation & Conservation Special Publication No. 12. Social cognition parallax interview corpus (SCOPIC)*. 1–21. Honolulu: University of Hawai’i Press. (<http://hdl.handle.net/10125/24742>)
- Barth, Danielle & Stefan Schnell. 2021. *Understanding corpus linguistics*. London: Routledge. doi:10.4324/9780429269035
- Bell, Allen. 1984. Language style as audience design. *Language in Society* 13(2). 145–204. doi:10.1017/s004740450001037x
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511621024.004
- Biber, Douglas. 1994. An analytical framework for register studies. In Biber, Douglas & Edward Finegan (eds.), *Sociolinguistic Perspectives on Register*. 31–56. Oxford: Oxford University Press.
- Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37. doi:10.1515/cllt-2012-0002
- Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge: Cambridge University Press. doi.org:10.1017/9781108686136

- Biber, Douglas, Jesse Egbert, Daniel Keller, & Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics* 171. 20–35. doi:10.1016/j.pragma.2020.09.018
- Biber, Douglas & Edward Finegan. 1994. Register and social dialect variation: An integrated approach. In Biber, Douglas & Edward Finegan (eds.), *Sociolinguistic perspectives on register*. 315–350. Oxford: Oxford University Press.
- Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9(1). 2–20. doi:10.1016/j.jeap.2010.01.001
- Brown, Penelope, & Steven C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511813085
- Brugman, Claudia, & Thomas J. Conners. 2019. Distinguishing properties of SMS and Twitter in Indonesian: A contrastive study. *Digital Scholarship in the Humanities* 34. 244–260. doi:10.1093/llc/fqy028
- Cardinaletti, Anna. 1990. Subject/object asymmetries in German null-topic constructions and the status of specCP. In Mascará, Joan & Marina Nespor (eds.), *Grammar in progress*. 75–84. Berlin: De Gruyter Mouton. doi:10.1515/9783110867848.75
- Clark, Herbert H. 1996. Common ground. In *Using language*. 92–122. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620539.005
- D’Arcy, Alexandra, & Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3). 255–285. doi:10.1017/S0954394515000101
- Davies, Mark. 2004. British National Corpus (from Oxford University Press). Oxford: Oxford University Press. (<https://www.english-corpora.org/bnc/>) (Accessed 2022-11-06.)
- Davies, Mark. 2007. The Corpus of Contemporary American English (COCA). (<https://www.english-corpora.org/coca/>) (Accessed 2022-11-06.)
- Duden – Die deutsche Rechtschreibung. 2020. vol. 1, 28th edn. 2020. Mannheim: Bibliographisches Institut & F.A. Brockhaus AG.
- ELAN [Computer software]. 2022. The Language Archive, Nijmegen: Max Planck Institute for Psycholinguistics. (<https://archive.mpi.nl/tla/elan>) (Accessed 2022-04-01.)
- Ferguson, Charles A. 1959. Diglossia. *Word*. 15, 325–340.
- Ganzeboom, Harry B. G. & Donald J. Treiman. 1996. Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research* 25(3). 201–239. doi:10.1006/ssre.1996.0010
- Ghafar Samar, Reza & Tej Bhatia. 2017. The future of “future”: A persian perspective on grammaticalization of future marking. *Asia-Pacific Language Variation* 3(2). 130–159. doi:10.1075/aplv.16011.gha
- Goebel, Zane. 2005. An ethnographic study of code choice in two neighbourhoods of Indonesia. *Australian Journal of Linguistics* 25. 85–107. doi:10.1080/07268600500113674
- Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3). 471–496. doi:10.1017/s1360674314000136
- Gries, Stefan T. & John Newman. 2014. Creating and using corpora. In Podesva, Robert J & Devyani Sharma (eds.), *Research Methods in Linguistics*. 257–287. Cambridge: Cambridge University Press. doi:10.1017/cbo9781139013734.015
- Haerpfher, Christian, Robert Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, et al. (eds.) 2022. *World Values Survey: Round Seven – Country-Pooled Datafile*. (version 5.0). Madrid, Spain. JD Systems Institute. doi:10.14281/18241.20
- Haig, Geoffrey & Shirin Adibifar. 2019. Referential Null Subjects (RNS) in colloquial spoken Persian: Does speaker familiarity have an impact? In Korangy, Alireza & Behrooz Mahmoodi-Bakhtiari (eds.), *Essays on Typology of Iranian Languages*. 102–121. Berlin: De Gruyter Mouton. doi:10.1515/9783110604443-006

- Haig, Geoffrey & Stefan Schnell (eds.). 2022. Multi-CAST: Multilingual corpus of annotated spoken texts (version 2311). Bamberg: University of Bamberg. (multicast.aspra.uni-bamberg.de/) (Accessed 2022-01-10.)
- Halliday, Michael A. K. 1978. *Language as social semiotic : The social interpretation of language and meaning*. London: Edward Arnold.
- Halliday, Michael A. K., & Ruqaiya Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
- Hasan, Ruqaiya. 2014. Towards a paradigmatic description of context: Systems, metafunctions, and semantics. *Functional Linguistics* 1(1). 9. doi:10.1186/s40554-014-0009-y
- Hellwig, Birgit. 2019. Linguistic diversity, language documentation and psycholinguistics: The role of stimuli. In Lahaussais, Aimée & Marine Vuillermet (eds.), *Methodological tools for linguistic description and typology*. Honolulu: University of Hawai'i Press. (<http://hdl.handle.net/10125/24855>)
- Himmelman, Nikolaus. P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196. doi:10.1515/ling.1998.36.1.161
- Himmelman, Nikolaus. P. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelman, & Ulrike Mosel (eds.), *Essentials of language documentation*. 1–30. Berlin: De Gruyter Mouton. doi:10.1515/9783110197730.1
- Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Seifart, Frank, Geoffrey Haig, Nikolaus P. Himmelman, Dagmar Jung, Anna Margetts, & Paul Trilsbeek (eds.), *Potentials of language documentation: Methods, analyses, and utilization*. 111–117. Honolulu: University of Hawai'i Press. (<http://hdl.handle.net/10125/4523>)
- IDS. 2023. Datenbank für Gesprochenes Deutsch (DGD). Mannheim: Leibniz-Institut für deutsche Sprache. (<http://dgd.ids-mannheim.de>) (Accessed 2023-01-10.)
- Inglehart, Robert & Christian Welzel. 2005. *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press. doi:10.1017/CBO9780511790881
- International Labour Office. 2012. International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables. Genf: International Labour Office.
- Kilgarriff, Adam & Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3). 333–347. doi:10.1162/089120103322711569
- Krause, Thomas & Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1). 118–139. doi:10.1093/llc/fqw057
- Krifka, Manfred. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55(3-4). 243–276. doi:10.1556/aling.55.2008.3-4.2
- Labov, William. 2006. *The social stratification of English in New York City*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511618208
- Lehmann, Nico. 2024. *The intricacies of register variation across languages*. PhD thesis. Humboldt-Universität zu Berlin: Berlin.
- Lüdeling, Anke, Artemis Alexiadou, Aria Adli, Karin Donhauser, Malte Dreyer, Markus Egg, Anna Helene Feulner, Natalia Gagarina, et al. 2022. Register: Language users' knowledge of situational-functional variation: Frame text of the first phase proposal for the CRC 1412. *Register Aspects of Language in Situation* 1(1). 1–59. doi:10.18452/24901
- Lüdeling, Anke & Merja Kytö. 2009. *Corpus linguistics: An international handbook*. Berlin: De Gruyter Mouton.
- Lüpke, Friederike. 2010. Research methods in language documentation. *Language Documentation and Description* 7. 55–104. doi:10.25894/ldd227
- Marneffe, Marie-Catherine de, Christopher Manning, Joakim Nivre, Daniel Zeman (2021): *Universal Dependencies*. In: *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.

- McEnery, Tony & Richard Xiao. 2007. Chapter 2. Parallel and comparable corpora: What is happening? In Anderman, Guinlla & Margaret Rogers (eds.), *Incorporating corpora: The linguist and the translator*. Bristol: Multilingual Matters. doi:10.21832/9781853599873-005
- McEnery, Tony, Richard Xiao, & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge. (<https://katalog.ub.uni-leipzig.de/Record/0-510572618>) (Accessed 2022-07-08.)
- Modaressi-Tehrani, Yahya. 1978. *A sociolinguistic analysis of Modern Persian*. Lawrence: University of Kansas.
- Neumann, Stella. 2014. *Contrastive register variation: A quantitative approach to the comparison of English and German*. Berlin: De Gruyter Mouton.
- Nivre, Joakim., Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, et al. 2020. Universal Dependencies v2: An evergrowing multilingual Treebank collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 4034–4043. Marseille, France: European Language Resources Association. (<https://aclanthology.org/2020.lrec-1.497>) (Accessed 2022-04-01.)
- Nurani, Luisa Marliana. 2015. *Changing language loyalty and identity: An ethnographic inquiry into the societal transformation of the Javanese people* (PhD thesis). Arizona State University.
- OECD. 2015. International Standard Classification of Education (ISCED) 2011 overview. In *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*. Paris: OECD Publishing. doi:10.1787/9789264228368-3-en
- Pescuma, Valentina N., Dina Serova, Julia Lukassek, Antje Sauermann, Roland Schäfer, Aria Adli, Felix Bildhauer, Markus Egg, et al. 2022. Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. *Frontiers in Psychology*. doi:10.3389/fpsyg.2022.964658
- Pfeiler, Barbara & Lenka Zámešková. 2006. Bilingual education: Strategy for language maintenance or shift of Yucatec Maya? In Hidalgo, Margarita (ed.), *Mexican indigenous languages at the dawn of the twenty-first century*. 294–313. Berlin: de Gruyter Mouton. doi:10.1515/9783110197679.3.294
- Poedjosoedarmo, Soepomo. 1968. Javanese speech levels. *Indonesia* 6. 54–87.
- Poynton, Cate. 1985. *Language and gender: Making the difference*. Geelong, Australia: Deakin University.
- Ravid, Dorit & Liliana Tolchinsky. 2002. Developing linguistic literacy: A comprehensive model. *Journal of Child Language* 29(2). 417–447. doi:10.1017/S0305000902005111
- Robson, Stuart & Singgih Wibisono. 2002. *Javanese-English Dictionary*. Singapore: Periplus Editions.
- Sato, Yosuke. 2015. Argument ellipsis in Javanese and voice agreement. *Studia Linguistica* 69. 58–85. doi:10.1111/stul.12029
- Sato, Yosuke & Simin Karimi. 2016. Subject-object asymmetries in Persian argument ellipsis and the anti-agreement theory. *Glossa: A Journal of General Linguistics* 1(1). 8. doi:10.5334/gjgl.60
- Schäfer, Lisa. 2021. Topic drop in German: Empirical support for an information-theoretic account to a long-known omission phenomenon. *Zeitschrift Für Sprachwissenschaft* 40(2). 161–197. doi:10.1515/zfs-2021-2024
- Schnell, Stefan & Danielle Barth. 2018. Discourse motivations for pronominal and zero objects across registers in Vera'a. *Language Variation and Change* 30(1). 51–81. doi:10.1017/s0954394518000054
- Seifart, Frank, Ludger Paschen, & Matthew Stave. 2022. *Language Documentation Reference Corpus (DoReCo) 1.2*. Berlin: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). doi:10.34847/nkl.7cbfq779
- Sneddon, J. 2003a. Diglossia in Indonesian. *Bijdragen Tot de Taal-, Land- En Volkenkunde* 159(4). 519–549.
- Sneddon, J. 2003b. *The Indonesian language: Its history and role in modern society*. Sydney: UNSW Press.

- Stevanovic, Melisa, & Anssi Peräkylä. 2012. Deontic authority in interaction: The right to announce, propose, and decide. *Research on Language and Social Interaction* 45(3). 297–321. doi:10.1080/08351813.2012.699260
- Straka, Milan, Jan Hajic, & Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris: ELRA (European Language Resources Association).
- Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1). doi:10.1075/rs.18006.szm
- Szmrecsanyi, Benedikt & Alexandra Engel. 2023. A variationist perspective on the comparative complexity of four registers at the intersection of mode and formality. *Corpus Linguistics and Linguistic Theory* 19(1). 79–113. doi:10.1515/cllt-2022-0031
- Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin: De Gruyter Mouton. doi:10.1515/9783110896541
- Verhoeven, Elisabeth & Nico Lehmann. 2018. Self-embedding and complexity in oral registers. *Glossa: A Journal of General Linguistics* 3(1). 1–30. doi:10.5334/gjgl.592
- Wiese, Heike, Artemis Alexiadou, Shanley Allen, Oliver Bunk, Natalia Gagarina, Kateryna Iefremenko, Esther Jahns, Martin Klotz et al. 2019a. RUEG Corpus 0.2.0. Zenodo. doi:10.5281/zenodo.3236069
- Wiese, Heike, Artemis Alexiadou, Shanley Allen, Oliver Bunk, Natalia Gagarina, Kateryna Iefremenko, Esther Jahns, Martin Klotz et al. 2019b. RUEG Corpus Documentation. Berlin: Humboldt-Universität zu Berlin. (<https://korpling.german.hu-berlin.de/rueg-docs/v0.2/annotations.html>) (Accessed 2021-05-01)
- Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description* 12. 19–36. doi:10.25894/ldd161

Appendix

Table 2: A day in the life of a data collection for the Lang*Reg protocol. An exemplary routine describing the complete procedure of the data collection on a single day with four participants in two groups (A and B). We had a researcher or a local data collection collaborator as well as two local assistants for the recordings to manage all the events. The timings may differ depending on punctuality of participants, number of assistants helping the researcher and distance between recording locations. All in all, this procedure takes about 4-5 hours for one group.


Time	Duration (min)	Location	Group	Event
10:00	15	Private room	A	Arrival of first group
				The researcher and assistants greet the first group and bring them to the first recording location. Participants are given a general introduction to the course of events for the next few hours as well as an information sheet and a form of consent to sign. They are told that they will engage in a series of conversations which will be recorded by audio devices. They should talk freely during the conversation and ignore the recording device.
10:15	5	Private room	A	Preparing first recording
				Participants are seated comfortably in the recording location for private sessions and are offered something to drink and snacks. The researcher hands them the microphones, which they attach to their clothes. The transmitter is placed out of sight, for example in pockets. They are instructed about the tasks: first one of them tells a story about travelling for two minutes, then they hear a knock and switch – then the second person tells the story. After two minutes, they hear another knock and have 15 minutes for a casual conversation, in which they can talk about travelling in general, but are free to talk about whatever comes up as long as they keep talking.
10:20	20	Private room	A	Start of first recording: Situation 2-3
				The researcher presses record and leaves the room, closing the door.
10:30	15	Private room	B	Arrival of second group
				An assistant greets the second group and brings them to the recording location. They are placed a short distance away, out of sight, from the recording room and receive the same introduction as group A.
10:40	–	Private room	A	End of first recording
				The researcher signals to stop the conversation. They then give the instructions that one participant will stay in the room while the other has a waiting time of about 15 minutes. The one staying is told that another person will come in soon and that they should talk for 15 minutes about travelling etc. Ideally, they would find out something about the experiences or preferences of the other person when it comes to travelling. The one that will wait is shown a place away from the recording room, out of sight and away from group B, to rest. It is important that the two strangers do not meet, or even see each other, until the conversation begins, so this is why the room is out of sight.
10:40	–	Private room	B	Fetching group B
				One of the participants from group B is fetched by the researcher. They receive the microphone, which they attach to their clothes, and the transmitter, which they hide in a pocket or bag. The researcher brings them to the front of the recording room where they receive the instructions. They are to enter the room and sit down comfortably. There will be drinks and snacks. Another participant is waiting inside. They can talk about travelling for about 15 minutes but can also talk about anything else that comes up. They might try to get an idea of what the other participant has experienced or prefers when it comes to travelling. After 15 minutes they will hear a knock on the door, which is when the recording finishes.
10:45	15	Private room	A+B	Start of second recording: Situation 4
				The researcher presses start on the recorder before the participant from group B enters. Participant from group B enters.

Time	Duration (min)	Location	Group	Event
11:00	–	Private room	A+B	End of second recording: Situation 4
				The researcher knocks and turns off the recorder. The participant from group A is led to their friend by an assistant. The participant from group B is led to their friend by the researcher.
11:05	–	Private room	A+B	Fetching second pair
				The researcher leads the second participant of group B to the recording room and gives the instructions as before. They may talk about travelling, inquire about the others experiences and preferences, and are free to talk about what they want. The other person will enter soon. An assistant has brought the second participant from group A and given the same instructions.
11:05	15	Private room	A+B	Start of third recording: Situation 4
				The researcher presses start on the recorder before the participant from group A enters. Participant from group A enters.
11:20	–	Private room	A+B	End of third recording: Situation 4
				The researcher knocks and turns off the recorder. The participant from group A is led to their friend by an assistant. The participant from group B stays in the room.
11:20	–	Private room	B	Fetching group B participant
				The second participant from group B is fetched by an assistant and is seated in the recording room. They receive the microphone and transmitter. Group B is instructed as follows: first one of them tells a story about travelling for two minutes, then they hear a knock and switch – then the second person tells the story. After two minutes, they hear another knock and have 15 minutes for a casual conversation, in which they can talk about travelling in general, but are free to talk about whatever comes up as long as they keep talking.
11:25	20	Private room	B	Start of fourth recording: Situation 2-3
				The researcher presses record and leaves the room, closing the door.
11:25	45-60	Private room	A	Social questionnaire
				An assistant instructs group A about the social questionnaire, handing them each an iPad or computer on which they follow the LimeSurvey instructions. At least one assistant stays with group A for questions arising from the questionnaire.
11:50	–	Private room	B	End of fourth recording: Situation 2-3
				The researcher knocks and turns off the recorder. Both participants of group B are now lead outside by the researcher.
11:50	5	Taxi	B	Arrival of the taxi driver
				Outside, the taxi driver is arriving with his taxi. The researcher greets the taxi driver and helps him attach the microphone, placing the transmitter in the middle console. The taxi driver signs the consent forms and is reminded that they are to drive around for at least 15 minutes and talk to the participant about travelling or anything that comes up. The taxi driver has been instructed prior to the recording about ways to bring the conversation naturally towards travelling.
11:55	15	Taxi	B	Start of fifth recording: Situation 5
				One participant is told to enter in the back with a microphone and transmitter. They are to drive in the taxi for 15 minutes and talk to the taxi driver. The researcher presses record and leaves the recorder on the seat next to the participant. They close the door and the taxi driver drives away.
12:10	–	Taxi	B	End of fifth recording: Situation 5
				The taxi driver returns and the researcher opens the door and stops the recording. The participant gets out of the taxi and gives the microphone to the friend.

Time	Duration (min)	Location	Group	Event
12:10	15	Taxi	B	Start of sixth recording: Situation 5
				The other participant enters the car and the researcher presses play and closes the door.
12:25	–	Taxi	A	Fetching group A
				Group A has finished the social questionnaire and had a short break. The assistant leads them to the taxi driver location, where they wait together with the researcher and one participant of group B. Group A is instructed that they will enter a taxi and talk to the taxi driver for 15 minutes next.
12:25	–	Taxi	A+B	End of sixth recording: Situation 5
				The taxi driver returns and the researcher opens the door and stops the recording. The participant gets out of the taxi and gives the microphone to the first participant of group A.
12:30	15	Taxi	A	Start of seventh recording: Situation 5
				The first participant from group A enters the car and the researcher presses play and closes the door.
12:30	45-60	Private room	B	Social questionnaire
				An assistant instructs group B about the social questionnaire, handing them each an iPad or computer on which they follow the LimeSurvey instructions. At least one assistant stays with group B for questions arising from the questionnaire.
12:45	–	Taxi	A	End of seventh recording: Situation 5
				The taxi driver returns and the researcher opens the door and stops the recording. The participant gets out of the taxi and gives the microphone to the friend.
12:45	15	Taxi	A	Start of eighth recording: Situation 5
				The other participant enters the car and the researcher presses play and closes the door.
1:00	–	Taxi	A	End of eighth recording: Situation 5
				The taxi driver returns and the researcher opens the door and stops the recording.
1:00	10	Taxi	A	Farewell taxi driver
				The researcher thanks the taxi driver.
1:10	20	Taxi	A	Walk to professor's office
				The researcher and group A move to the last recording location, the professor's office.
1:30	5	Office	A	Arrival at professor's office
				The researcher greets the professor while group A waits somewhere near. The professor signs the consent forms, attaches the microphone and is instructed again to talk with each participant for 15 minutes until they hear a knock. The professor should try to steer the conversation towards moving around in times of the pandemic.
1:30	–	Private room	B	Break
				After finishing the social questionnaire, group B has a break until about 2:00. The assistant shows them how to get to the last recording location, the professor's office.
1:35	15	Office	A	Start of ninth recording: Situation 6
				The first participant of group A enters the professor's office with attached microphone, which has already been turned on by the researcher. They were instructed to talk with a professor in the office for 15 minutes.
1:50	–	Office	A	End of ninth recording: Situation 6

Time	Duration (min)	Location	Group	Event
				The researcher knocks on the door and stops the recording. The participant exits the room and gives the microphone to their friend.
1:50	30-60	Office	A	Start of writing assignment: Situation 1
				An assistant takes the participant of group A to a secluded area (or alternatively back to the private room if close by) where they can record situation 1, the writing assignment. They receive pen and paper and are instructed to write the story they have told to their friend as a letter to the same friend as if they hadn't told it to the friend before.
1:50	15	Office	A	Start of tenth recording: Situation 6
				The second participant of group A enters the professor's office with attached microphone, which has already been turned on by the researcher. They were instructed to talk with a professor in the office for 15 minutes.
2:00	–	Office	B	Arrival of group B at professor's office
				Group B arrives at the professor's office and may wait near the office (away from group A to prevent exchanges about the situation).
2:05	–	Office	A	End of tenth recording: Situation 6
				The researcher knocks on the door and stops the recording. The participant exits the room and gives the microphone to the researcher.
2:05	30-60	Office	A	Start of writing assignment: Situation 1
				An assistant takes the second participant of group A to a secluded area (or alternatively back to the private room if close by) where they can record situation 1, the writing assignment. They receive pen and paper and are instructed to write the story they have told to their friend as a letter to the same friend as if they hadn't told it to the friend before.
2:10	15	Office	B	Start of eleventh recording: Situation 6
				The first participant of group B enters the professor's office with attached microphone, which has already been turned on by the researcher. They were instructed to talk with a professor in the office for 15 minutes.
2:25	–	Office	B	End of eleventh recording: Situation 6
				The researcher knocks on the door and stops the recording. The participant exits the room and gives the microphone to their friend.
2:25	30-60	Office	B	Start of writing assignment: Situation 1
				An assistant takes the participant of group B to a secluded area (or alternatively back to the private room if close by) where they can record situation 1, the writing assignment. They receive pen and paper and are instructed to write the story they have told to their friend as a letter to the same friend as if they hadn't told it to the friend before.
2:25	15	Office	B	Start of eleventh recording: Situation 6
				The second participant of group B enters the professor's office with attached microphone, which has already been turned on by the researcher. They were instructed to chat with a professor in the office for 15 minutes.
2:45	–	Office	B	End of eleventh recording: Situation 6
				The researcher knocks on the door and stops the recording. The participant exits the room and gives back the microphone.
2:45	10	Office	B	Farewell to professor
				The researcher thanks the professor.

Time	Duration (min)	Location	Group	Event
2:45	30-60	Office	B	Start of writing assignment: Situation 1
				An assistant takes the second participant of group B to a secluded area (or alternatively back to the private room if close by) where they can record situation 1, the writing assignment. They receive pen and paper and are instructed to write the story they have told to their friend as a letter to the same friend as if they hadn't told it to the friend before.
3:00	5	Office	A	Post-experiment survey
				Once group A has finished the writing assignment, the researcher asks them to fill out the quick post-experiment survey.
3:10	–	Office	A	Farewell group A
				The research team thanks group A.
3:15	5	Office	B	Post-experiment survey
				Once group B has finished the writing assignment, the researcher asks them to fill out the quick post-experiment survey.
3:20	–	Office	B	Farewell group B
				The research team thanks group B.


Nico Lehmann
 nico.lehmann@hu-berlin.de
 orcid.org/0000-0002-5532-8244

Vahid Mortezaour
 orcid.org/0000-0002-6832-0057

Jozina Vander Klok
 orcid.org/0000-0001-8550-3181

Zahra Farokhnejad
 orcid.org/0009-0007-0837-8824

David Müller
 orcid.org/0009-0003-2431-6675

Elisabeth Verhoeven
 orcid.org/0000-0001-8775-8567

Aria Adli
 orcid.org/0000-0002-2177-7375