Chapter 3

Mapping the mappings and then containing them all: Quality assurance, interface modeling, and epistemology in complex corpus projects

Anna Shadrova[a], Martin Klotz[a], Rahel Gajaneh Hartz[a] & Anke Lüdeling[a]
[a]Humboldt-Universität zu Berlin

Building on our experience with the structured and deeply annotated RUEG corpus, this contribution summarizes methodological aspects of corpus modeling, compilation and analysis. We argue that corpus data – as opposed to experimental data – is characterized primarily by its open-endedness, and thus corpus-linguistic work is characterized primarily by the task of structuring and ontologizing data on a range of linguistic levels based on a range of linguistic models and with the aid of a range of technical models. As the unification of these diverse models requires interface definitions, it follows that annotation, modeling, analysis, and the epistemological embedding of categories and analytical processes must go hand in hand. For any large-scale corpus project, this complexity requires a division of labor that creates distributed and decentralized knowledge, which we understand as more interfaces requiring definitions. We conclude that the epistemic gains that can be procured from corpus projects go beyond the actual corpus resource and beyond discussions of technical representations, and that it would benefit the field to synthesize this type of modeling knowledge more productively for a positive impact on future projects as well as a better understanding of the emergent properties of situated language production.

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

1 Introduction

This paper summarizes aspects of the corpus-methodological research that took place within the Research Unit *Emerging Grammars* (RUEG) in 2018–2024 and the compilation of the corpus it created from elicited data, which we will refer to as the RUEG corpus (Lüdeling et al. 2024, Klotz et al. 2024). RUEG collected and analyzed data from heritage speakers of Greek, Russian, and Turkish in Germany and the US, German in the US, and Kurmanji (Kurdish) in Turkey, with all participants elicited in both of their languages; and from monolingually-raised participants in the countries where those heritage languages represent the majority language (e.g. Turkish in Turkey, Greek in Greece etc.). This was done with the purpose of comparing emergent patterns on various linguistic levels fostered by the speaker-internal language contact situation in the minds of multilingual individuals, and in correlation with situational aspects such as formality and modality. While challenges and constraints occur as concrete and specific in any research project, many challenges we have encountered arise from a generalizable requirement of mapping models and interfaces. We believe this requirement should receive more attention for the benefit of future corpus-related work. This paper thus is not primarily concerned with a description of the corpus, neither is it about technical standards of corpus compilation or data architectures, or about specific linguistic phenomena.[1] Rather, it presents a conceptual reflection of the complexity and interdependence of modeling decisions that this type of research requires and that have not found much attention in the debate until now.[2]

RUEG and its corpus are complex units created in a strongly collaborative effort by contributors from a variety of linguistic subfields and levels of experience. Combining data in six languages elicited in five countries from over twenty speaker cohorts,[3] and in recording spoken and written data based on carefully

[1] For a more in-depth discussion of either of those aspects, please view Wiese (2020), the continuously updated corpus documentation under https://korpling.german.hu-berlin.de/rueg-docs/latest/, and the other chapters in this volume.

[2] As a reflection of this complexity and interdependence as well as the current lack of methodological and meta-theoretical integration, a certain degree of oscillation between the concrete examples from the RUEG process and corpus and the abstraction of principles of corpus modeling could not be avoided. This may appeal more or less in different subsections to readers from different specializations, e.g. annotators, technicians, or linguistic experts. Readers are encouraged to skip ahead from subsections that provide too many low-level details for their interest.

[3] Adolescent/adult monolingual/bilingual speakers in each country plus some trilinguals, see more in Section 2.

prepared prompt material, the RUEG corpus is one of the most complexly designed corpus projects to exist to date. This is a reflection of the RUEG project's research goals, as it brings together a speaker-centric with a variety- or linguacentric perspective on language data (describing speaker behavior vs. language patterns), while also attempting a comparison of phenomena on various linguistic levels which are not perfectly identical between the chosen languages. It further compares linguistic structures in different modalities and degrees of formality and aims to find and map emergent patterns in multilingual contexts, i.e., patterns that are influenced by language contact and individual variability, and that may have not yet been fully linguistically described for the respective languages. Moreover, RUEG intends both quantitative and a qualitative descriptions and thus also interfaces with diverse analytical frameworks.

As visualized in Figure 1, the corpus compilation process, the project structure and the data architecture reflect the complexity of RUEG's research agenda. This is partially due to the phenomena of interest, which require deep linguistic annotation that cannot be performed automatically at adequate quality. Data therefore has to be annotated manually, collaboratively, and additively. In this, some layers of annotation and pre-processing depend on others, therefore requiring an iterative workflow and many decisions around the adequacy and usefulness of the chosen categorization, i.e., the linguistic model, at each level.

In this contribution, we argue that the central challenge of the compilation, infrastructure provision, maintenance, and analysis of task-based corpora like the RUEG corpus lies in the mapping of different models and their interfaces. "Modeling" here refers to the process of explicitly defining the object or process of interest in terms of a selected theoretical, conceptual, methodological, or other tradition or framework (which are also models themselves).[4] "Interface mapping" refers to the process of identifying and defining the connective properties and elements between two or more of those modeled objects. We will discuss these notions in greater depth in Section 3.

With this, we offer a perspective on corpora that does not primarily center the result of their compilation, namely the corpus itself as a reusable resource, but one that views the process of thought and construction with the interdependence of all steps in its own epistemic right. The corpus resource is then just a view of

[4] We are aware that different disciplines and areas understand modeling in different ways and that this is only one of the possible meanings of modeling (see for example Stachowiak 1973, White 1973) and more generally the debates in the JLM (https://jlm.ipipan.waw.pl/index.php/JLM). We will not be able to discuss other approaches and philosophies of modeling here, and will instead provide definitions of our understanding where necessary.

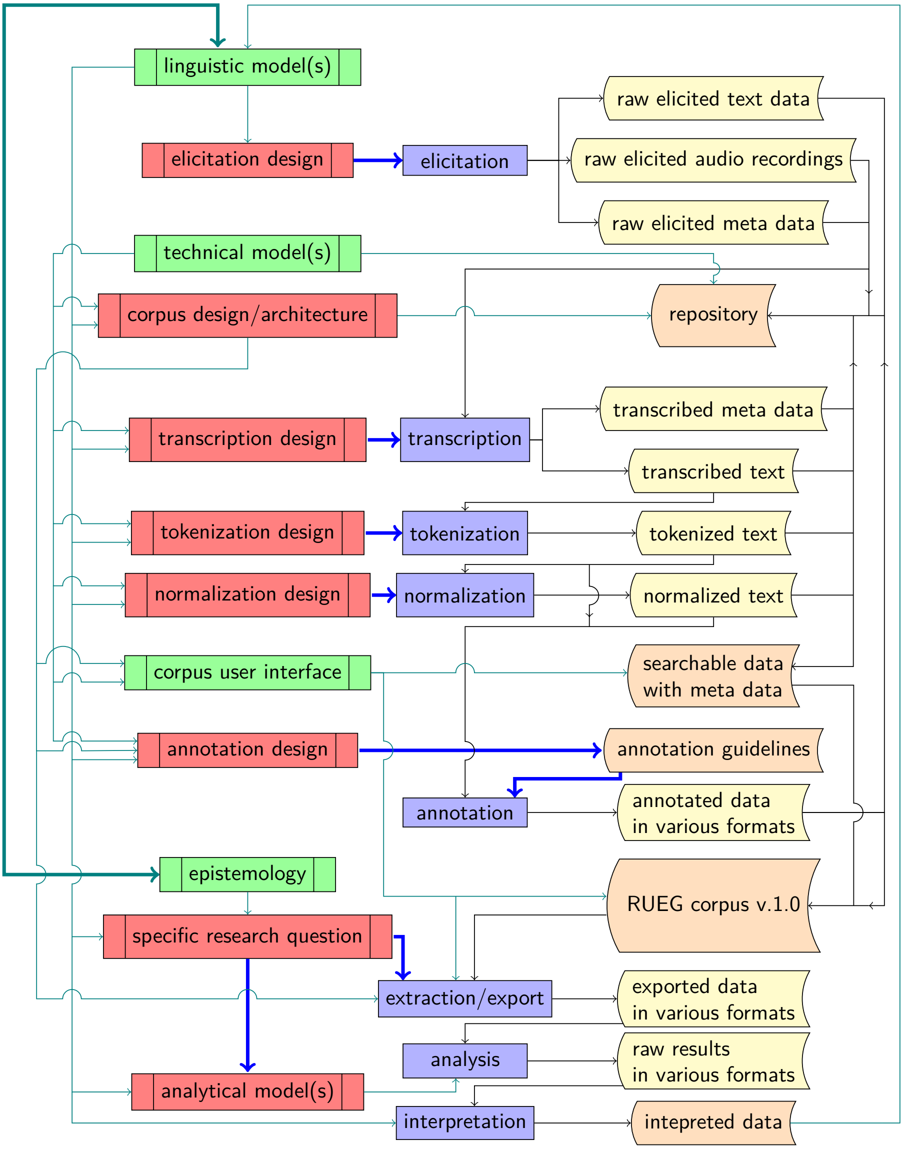

Figure 1: Processes involved in the compilation of the RUEG corpus. Arrow colors signify different types of relations: teal – *influences*, *limits*, or *defines access to*; blue – *defines*; black – *results in* or *is passed over to*. Colors and shapes signify type of process, model, or framework: boxes: green – underlying, project-independent models or infrastructures; red – design choices made within the project; violet – processes performed by project members; book-shaped: yellow – process products; orange – stable, reusable products.

one stage in its development. Since the RUEG corpus was developed under significant time constraints relative to its complexity, it also provides insight into the challenge of handling subprocesses in parallel in spite of their mutual interdependence. This simultaneity poses specific demands on the technical realization as well as high interactional and communicative pressure on the collaborators, especially since different types of models, processes, and objects (marked in different colors in Figure 1) are often handled by different project members. In RUEG, this is further complicated by the groups' decentralized organisation across several locations in Germany.

The text is structured as follows. We will first provide an overview of the corpus design, the elicited material, and the core research questions in Section 2. We will then introduce the notion of linguistic modeling and discuss how different models interlock in an interdisciplinary project and why that poses a specific problem in corpus data (Section 3), followed by a discussion of some of the modeling processes and common challenges in Section 4. Specifically, we will discuss differences between sequential (what happens in time) and structural aspects (which choices depend on previous ones) of the model; how to map diverse and partially incompatible models; and how to map models that are not yet fully defined. Wherever useful, we will illustrate our points with examples from the RUEG corpus and context, although we believe that the problems are more structural and relevant to all of corpus linguistics and especially projects building and analyzing task-based corpora.

We offer five central theses in this paper:

1. Since corpus data is not experimental in nature, most of the epistemological weight of any argument derived from corpus data lies in the post-elicitation modeling, selection, and analysis.

2. Modeling decisions are interface definitions connecting different models. Models exist on the technical side (such as data structures or database architectures) and on the linguistic side (such as different linguistic models projected onto the data through annotation). In a collaborative environment, each collaborator's understanding is also a (usually implicit) model. With a distribution of labor between e.g. project managers, annotators, technical experts, statistical analysts, and so on, corpus projects create distributed knowledge, which necessitates interface definitions between collaborators in order to remain consistent.

3. Interface definitions are mappings of the categories of different models onto one another and definitions of their potential interactions. They are

not usually obvious or available a priori for sufficiently diverse models, e.g. technical infrastructure and linguistic content, or linguistic theory at the lexical vs. the syntactic level. The number of interfaces is at least a quadratic function of the number of collaborators, which means that raising the number of collaborators shifts the balance of a corpus project towards more required mapping, i.e., communicative overhead. It is questionable whether these communicative demands can be satisfied within the constraints of common research infrastructure for very large corpus projects.

4. Linguistic depth is therefore most approachable through small to medium-sized, well-controlled, and manually annotated corpora and within infrastructures that allow for as many of the collaborators as possible to participate in communication and analytical decision-making at each step.

5. The corpus and scientific publications based on its data are not a comprehensive representation of the set of epistemes that arise from corpus projects. Rather, procedural aspects, including linguistic and technical modeling decisions, represent knowledge in their own epistemic right. This type of knowledge cannot be passed on in full through continued access and maintenance of a corpus-database or through corpus-specific documentation. Rather, the field needs more reflection on generalizable corpus-methodological epistemes beyond the discussion of technical standards.

While task-based corpus elicitations limit the active potential of participant responses, they do not effectively limit the variability or complexity of the data. Rather, even within this limitation, variability remains massive and the emergent properties of language are omnipresent. Corpus linguistics has the unique benefit of modeling the interaction of the full emergent quality of situated language production, but can only do so under epistemological clarity and demarcation of all relevant categories and interrelations. This requires a strong focus on the distillation and synthesis of models that are applicable to the specific case from a large number of available ones, which involves a process of idiosyncratic expertise development. It cannot (helpfully) be rushed and should not be viewed solely in terms of publications or resources generated.

2 The RUEG corpus

Within the RUEG project, language contact phenomena are investigated from different viewpoints. One perspective is a focus on possible influences of different heritage languages on the same majority language. Another one is a focus on possible influences of different majority languages on the same heritage language. A list of all the projects, their research topics and their perspectives can be found in chapter (Wiese, Allen, et al. 2025 [this volume]).

A heritage language is a language acquired naturalistically during first language acquisition that is different from the majority language spoken by the surrounding society (Pascual y Cabo & Rothman 2012, Montrul 2015). Literature points out that differences can be observed in heritage speakers compared to monolingual speakers regarding the heritage language as well as the majority language (Rothman 2009, Polinsky 2018). Differences of this kind, though, are not undisputed (Wiese et al. 2022, Özsoy & Blum 2023). Definitions for the terms *heritage language*, *heritage speaker* and *monolingual speaker* are also provided in the introductory chapter (Wiese, Allen, et al. 2025 [this volume]).

The design of the RUEG corpus is driven by several general research questions investigated by multiple research groups from different cities in Germany in several overlapping phases of data elicitation, annotation, correction, and statistical analysis.

To obtain data appropriate for the research questions, an elicitation procedure was defined (Wiese 2020), for which participants are presented with a short stimulus video showing a series of events that lead to a minor traffic accident, which they are then asked to report in four different situations defined across two modes (spoken vs. written) and two degrees of formality (formal vs. informal). The setup acknowledges the sensitivity of language use to the production situation (in order to assess register knowledge, see e. g. Lüdeling et al. 2022). For differentiating findings specific for heritage language from general trends within a majority language of a country, speakers with no heritage language background were included as well.[5] To also address questions of speaker generation, elicitations spanned across two age groups, adolescents and adults. Additional knowledge on the participants' socioeconomic background, language biography and habits, and personality was collected with a digital questionnaire. Data collection took place over several months in five countries and for six languages (see Table 1).

[5] For simplicity, we will refer to this speaker cohort as monolinguals, although we are aware that most speakers are not truly monolingual in an urban and digitized environment.

Table 1: Languages elicited and country of elicitation by subcorpus

	Country				
Language	Germany	Greece	Russia	Turkey	USA
Core corpus					
German	✓				✓
Greek	✓	✓			✓
Russian	✓		✓		✓
Turkish	✓			✓	✓
English					✓
Extensions					
Kurmanji				✓	

We chose a multi-layer architecture (Zeldes 2018) because of its extensibility and flexibility with regard to number and types of annotation layers. Additionally, such an architecture allows to enrich the corpus with extensive metadata obtained from the participant questionnaire. The research questions require linguistic analysis of phenomena from different linguistic domains at different degrees of granularity. Spoken data is carefully transcribed, written data is automatically extracted and reformatted (Schmidt & Wörner 2014, Boersma 2001). Then, both types of data are passed on to further processing. We follow the tradition from historical linguistics to call the layer that is as close as possible to the original the *dipl*omatic layer. As a baseline for follow-up annotations, the dipl layer is then *norm*alized to canonicalized tokens, to which lexical and functional annotations (lemma, part of speech (PoS), morphological features), syntactic annotations (dependencies and constituency trees), or annotations regarding information status and referents are applied. A subgroup directly annotates prosodic features for the diplomatic tokens.

A centralized data repository is used to store the annotated data and their versions throughout the entire project. The repository is interfaced with automatic preprocessing pipelines for enriching the data with new annotations, testing it against several criteria of consistency and integrity, and to deploy a combined resource for search and visualization under a common technical model (Zipser & Romary 2010, Krause & Zeldes 2014, Krause & Klotz 2023). Also, the repository interface served as knowledge base for generating the future corpus documentation. It furthermore served for the coordination of the complex annotation process handled by multiple parties with different annotation tasks and scopes (for an illustration see Figure 1). The infrastructure around the RUEG corpus is designed to provide easy process for further iterations of annotation, documentation, investigation, and publication, i.e., linguistic research in general.

As previously mentioned, the RUEG corpus and its data are the outcome of a coordinated collaborative effort. Multiple research groups have annotated and continue to annotate the raw data in parallel, basing their work in specific expertise on different subfields of linguistics, which we consider the most interesting and challenging aspect of this endeavour. Annotations are designed and applied relative to each respective research question. For this, there are regular conceptual exchange meetings, after which annotation groups work mostly autonomously. Annotation processes are frequently executed by student assistants and PhD students in exchange, but not usually in direct collaboration, with postdoctoral researchers and professors. Annotation tasks can focus on a specific speaker group or a single language, but can also range across multiple languages. Depending on the type of annotation and its structure, different annotation tools are involved in an annotation workflow, which results in variable formats (representations).

3 Data, models, and interfaces

On the most abstract level, a corpus can be described as a collection of data. But what is data?

On the right side of Figure 1 we have visualized in orange a range of stable, reusable resources that have resulted from the work of the RUEG project. This most obviously includes the final 1.0 version of the RUEG corpus, but also its predecessors without manual annotations, a repository storing all materials and data in their various stages, annotation guidelines that can be reused for future research, and of course a number of publications that will serve the development of heritage language research and related fields. Typically, the corpus itself would be referred to as “data”, or as processed data, and the elicited material, the audio recordings, typed text, questionnaires and so on, as “raw data”.

Beyond this, RUEG has produced many intermediate stage products, visualized in yellow in Figure 1, which we will refer to as process products. This includes each of the transcribed, tokenized, and normalized text, and a range of annotation styles or tiers (e.g. syntactic, phonetic, or discourse-level), as well as data that was exported from the corpus and entered into other analytical processors, such as statistics scripts, and the results thereof. It is also common for researchers to regroup and filter data to their convenience outside of the provided corpus interface and add on-the-fly annotations during analysis. Data interpretation (visualization, etc.) may therefore be based on only a subset or an extension of the data actually provided in RUEG 1.0, and there may be many more versions of the RUEG corpus on individual researchers' computers that the authors are not aware of.

At each step of the process from the top left to the bottom right, "data" presents as a different type of object. In Figure 1, this can be read from the black arrows. The process of transcription, for instance, takes the raw elicited audio recordings and the raw elicited metadata, and yields transcribed metadata and transcribed text as a result. This *result* of one process is then passed on to the next process of tokenization and normalization *as data*. Similarly, the process of annotation takes *tokenized text data* that may or may not also be normalized, i.e., be the result of one or two processes, as data, and yields *annotated data* as a result. Finally, the researcher working on the final analysis will take the annotated data, or more generally the *corpus data* as an input to their analytical model. This results in new, e.g. quantitative *data* that is then interpreted.

At this point, the researcher is working with the result of a range of processes which each involve a number of linguistically relevant and often complexly interdependent choices. In Figure 1, these are summarized in the red and green boxes representing project-specific and external sets of views and choices respectively. These sets of views and choices are models.

A model in our sense is an abstract description of a phenomenon (an object or a process) that contains, i.e., defines, the elements and parameters relevant to the context of use and their interrelations. For example, in the RUEG context, the transcription design contains information on how to adequately transpose phonetic to graphematic information. But there is no information on how to adequately transcribe archaic handwriting to modern graphemes, because this is irrelevant in the RUEG context. The transcription design is a model, not a 1:1 representation of the process. It does not and cannot reflect or predict every single mapping from phonetic quality to graphematic representation. Rather, it is itself a product of a process of communication and mediation between several participants (frequently more experienced researchers instructing student assistants on how to transcribe, and transcribers bringing questions on unresolved cases). This process is informed by other models, for instance conventions of orthography, which in turn rely on etymology, language policy, etc.

In this way, all research is a reflection of a range of processes that typically result in a model (the result of a meeting; a paper; an annotation guideline; a corpus; a new theory; etc.) and are also informed by other models, which themselves are results of modeling processes, and so on.[6]

Any research process, and any type of data, in this way is explicitly or implicitly influenced by the models available to a field (paradigms, theories, hypotheses, explanations, etc.). This also means that data is never neutral, it is always a model that reflects underlying (presumed) ontologies.

All research further requires a definition of the relevant interfaces of all objects and processes involved. Any well-designed experiment, for instance, is defined in terms of how the expected or unexpected result would fit with the theoretical framework; and any function of a programming language is defined in terms of which data format it requires in order to be correctly executed. Interface definitions are hence simply descriptions of how the different elements (processes, objects, phenomena, etc.) fit with one another. Interface definitions are also models, i.e., functional abstractions of complex and fuzzy real-world phenomena into manageable sets of properties.

These processes are fundamental to all research, corpus-linguistic or otherwise. There are, however, aspects of corpus-linguistic research of the kind performed within the context of RUEG that pose specific challenges. These are, in no particular order:

a. the number of interfacing elements and processes – as Figure 1 shows, dozens of relations occur even in this synthesized and simplified description which does not reflect the iterations of subprocesses required to arrive at a higher-level concept such as *normalization*;

b. the path dependence or mutual interdependence of some of the processes involved (see Section 4.2);

c. the number of people involved, who each carry their own understanding of the material, i.e., their own model, effectively creating an interface with every person, every general and every project-specific model involved;

d. the open-ended character and fundamental abstractness of language;

e. the fact that corpora contain *naturalistic* language, which as an emergent phenomenon results in more than the sum of its parts through manifold and complex interrelations. This is independent of whether the context of utterance is also considered realistic or natural.

We will discuss the latter two points in detail below.

[6] Leonelli (2019) shows this very convincingly at the example of plant phenotyping in an interdisciplinary and digitized research setting.

3.1 The open-endedness of corpus data

Modern day linguistics frequently attempts a degree of objective categorization as it is traditionally sought in the natural sciences and computational subjects: linguistic phenomena are described to be *observed* in their context similar to observations of, say, the leaves of a plant.[7] However, unlike the leaves of a plant, we cannot touch, store away, cut, rotate, etc. language without a priori knowledge.[8] The multiplicity and variability of linguistic models in use shows that this can take very different shapes. The pluralicity of ways to view the elements of a language is a central issue of corpus linguistics, which at many stages is concerned precisely with the demarcation of linguistic units at different levels of abstraction and granularity. On a very basic level, this is the case in tokenization, i.e., the separation of character or audio strings into meaningful units (like phones, letters, words, punctuation signs, etc.).[9]

But it is equally the case in all annotation decisions (*does a certain string represent a case of category X?*, *in what shape does phenomenon Y occur in the data?*). Even for the classification of a seemingly delimited phenomenon like PoS, we find numerous models from various grammatical traditions, and several more from a more pragmatic computational perspective (for a discussion see e. g. Robins 1986, Knobloch & Schaeder 2000, Atwell 2008, Petrov et al. 2012).

Since there is no single, theory-neutral, "obviously natural" categorization, i.e., no clear-cut, extra-linguistically definable way of determining what constitutes a unit at any layer of linguistic granularity, a demarcation of meaningful units relative to a research question is of central importance to any corpus-based project. A project interested in discourse markers, for example, first has to decide what

[7] This is not to imply that the STEM fields always succeed in reaching objective conclusions, as debates around the replication crisis show. Recent research in biology also shows that results from the same dataset majorly depend on subjective choices even in seemingly objective analytical frameworks like inferential statistics (Gould et al. 2025).

[8] A biologist arriving on an island with previously unrecorded plants can go to a plant and touch, rotate, etc. its leaves, unless the leaves have some extremely unusual properties. A linguist hearing upon arrival on a remote island a previously unrecorded language has to put some effort in and meet informants to begin identifying even the basic building blocks of the unknown language, even if the language is not particularly unusual.

[9] While it may seem obvious where one word begins and another ends in many modern European languages, the process of tokenization actually requires many design choices regarding for instance the status of compounds, neologisms, the role of inflections, corrections and repairs. For languages that are not separated into orthographic words by convention (such as non-alphabetic or non-existent writing systems; agglutinative and polysynthetic languages; signed languages; historical languages prior to orthographic normalization; etc.), even more decisions have to be made (Bauer 2000, Schmid 2008). Even in European languages, we find phenomena like clitics or amalgamations that add complexity.

counts as a discourse marker, and then identify all occurrences in the data. This may seem obvious, but in our experience is not usually a priori defined in actual corpus projects and may not be a priori definable due to the inherent emergentist quality of the data.

Once this identification is performed in the form of annotation, the decisions, i.e., models, guiding said annotation will transcend through any further research that uses those annotations. A compiled and annotated corpus then will contain many modeling decisions upon which all further analysis depends, i. e., path-dependences in the data. Some categories may appear as though they fell into place naturally, but this only means that the categorization was performed based on implicit understanding, i. e., a hidden model. For annotators, this impression usually quickly dissipates when they are confronted with the full variability of their data. At this step, the general practice of dividing labor between annotators and model designers can become challenging. This is because, unless both parties participate in frequent in-depth discussion of the data, their implicit models of the data may easily diverge. It is not uncommon for those divergences to go unnoticed, only causing problems at later stages. Between annotators, this can be alleviated via agreement measures and guidelines, but when different models need to be fitted with one another, other strategies are required.

Why does corpus data give rise to divergent models so easily? We believe that this may be due to a mistaken perspective of *elicitation as quasi-experiment*: when analyzing corpus data, or when designing a task-based corpus, researchers often implicitly view it as a kind of experiment in which speakers are asked to perform task *x*, and the data that is extracted from the corpus is their response. However, that is not an accurate model. An experiment is the recording of a process in an a priori artificially constrained world, in which a participant can only respond within a predetermined response space (a response time; an eye movement pattern; a Likert scale; one out of a limited number of specified behaviors). The response may be multimodal (response time and Likert scale), but each response *space*, i.e., the number of options, is limited, hence also limiting the number of combinations. If a participant does not provide responses of the envisioned kind (e.g., if they glance away, or fall asleep and produce extremely long reaction times), they are excluded from the analysis. This also means that, in an experiment, the plausible kinds of analyses are embedded in the design. Experimental data on a Likert scale cannot be re-interpreted as eyetracking data instead. Rather, once the experiment is complete, the data exists in the expected format and can be passed on to the preconceived analytical framework. This means that in an experiment, the interface between the hypothesis/the theoretical model, the experiment design, and the outcome variable are all modeled during design, namely in constraining the response space in a finite and specific way.

This is not to say that all experimental data is unproblematic in its design or interpretation, or that experimental designs always succeed at isolating their factors as desired (Dewaele 2019, Gozli 2019). Rather, the point is that the data elicited from an experiment exists in a world closed and discrete by design, and hence the objects of interest for analysis and interpretation are known by design: as soon as the experimental data comes in, the results are determined. The data is delimited and speaks for itself (within the realms of the experiment – in experimental data, production roughly equals result with a deterministic transformation through a method of analysis).

In corpus data, we find more *qualitative* variability between responses, i.e., speakers responding to the task in cooperative, acceptable, and interpretable ways, but their responses lying outside our expectations in elicitation. Corpus data is continuous prior to analysis and open-ended (Moisl 2009). Continuous here refers to its quality of not naturally falling into categories, linguistic or otherwise. All demarcations of categories have to be performed post-hoc, including demarcation of words or sounds from the audio or string. This is even true of typed text, where the demarcation of words seems to be defined by the speaker, but will often be changed for normalization. It is open-ended in several ways:

a. Corpus data is not constrained to a limited response space – in responding to a task like the one prompted in RUEG, participants have a universe of choices at all linguistic levels.

b. Speakers make use of this freedom by responding in astoundingly variable ways, particularly in the lexical domain (Shadrova 2025). This shows even in the communicatively narrow space of the RUEG corpus, which means that a task-based limitation does not necessarily result in greater predictability of the lexical forms that will occur in the corpus.[10]

c. Beyond choosing different expressions for the same communicative purpose, speakers also frequently diverge in their interpretations of the situation, the task, or the communicative purpose and engage in playacting, change the genre, invent content, and so on.

d. Corpus data is not analytically deterministic and cannot *as such* be handed over to an analytical framework. It needs to always be filtered and rearranged before it becomes quantifiable.

[10] For instance, speakers in the German subcorpora find over 600 verbs to refer to the events and actions taking place in the 40 second video, see Shadrova et al. in preparation.

e. This process of filtering and rearranging itself is open-ended, as it depends on a multiplicity of analytical decisions and framework specifications.

f. Corpus data is also open-ended in a historical or dynamic sense, i.e., with respect to the emergence of structures that cannot be predicted because they are idiosyncratic, subject to ongoing language change, or as of yet undescribed. The latter is especially interesting in the RUEG context because the project aims at the identification of shifts in form and function, for which the models in use must be able to handle unknown forms or forms in new functions.

The complexity and open-endedness of the corpus linguistic analysis can hence be summarized as (1) massive variability and (2) interconnectivity with (3) an unpredictable number of productive interrelations between means of expressions and (4) concepts describing those means of expression (5) in a historically changing, dialectic, and non-deterministic space which (6) can be passed on to an open class of analytical models. In short, corpus data never speaks for itself, because it is a priori undelimited and analytically non-deterministic.

One may try to avoid the emergent space by filtering the data for a specific category. For instance, if interested in expressions of directionality, one might query for the occurrences of directional prepositional phrases or selected lexemes or morphemes, provided the corpus is equipped with the relevant annotations. But the full variability and complexity of how speakers express directionality (lexically/adverbially, constructionally, pragmatically, etc.) is *still present*, and raises the question of what should be considered in the conceptual and analytical model; it cannot be avoided in the linguistic discussion of the data. Filtering a corpus for a specific type of information does not undo the complexity and emergence that speakers bring to the data. At the very least, the analytical model has to specify the cases it excludes, adding more interrelations and descriptive complexity. In this way, corpus data is open-ended with respect to the factor combinations that emerge from it, because an unlimited number of demarcations and interpretations can be combined with one another, resulting in a perplexing combinatorial space even in small data.

The major challenge of corpus linguistics then is to delineate meaningful constructs in corpus data and to model their interrelations in an epistemologically convincing way relative to the linguistic and analytical frameworks employed. This makes corpus linguistics a *constructionist*, rather than an *analytical* field until rather late stages of the research process.[11] Only after all categories and in-

[11] This distinction might be reminiscent of the discussion about corpus-driven, theory-driven, or corpus-based approaches (Biber 2015). Framed in these terms, our approach may be labelled as both theory-driven and corpus-based.

terrelations have been constructed (e.g. annotated), can the analytical process of separating effects truly begin. It is of course this very work that provides insight into naturalistic language production in context, and frequently yields fascinating and surprising results. In order to do so, clear epistemologies and problem definitions are required. We will now present a range of ways in which this work was designed and performed in RUEG.

4 Models and interfaces in RUEG

We have mentioned that a complex project like RUEG consistently requires the mapping of various models into a common space. This process is defined by a number of factors, including the number of models to fit; their mutual compatibility, i.e., freedom of contradiction in representation, theoretical assumptions, or epistemologies; their interdependence; and their certainty, well-definedness or degree of ambiguity. Better understanding of each model contributes to ease of mapping. However, in reality, models are not always perfectly defined in every detail, and some happen to be partially incompatible. In the following sections, we will discuss some of the issues arising from the combinations of different types of models. One particularly interesting case in the RUEG context is the mapping of unknown models, i.e., the description of emergent phenomena which can only be described as a research outcome, rather than mapped onto the data a priori. Of course this is the most complex case, since it maps open-ended data to an open-ended modeling space at high uncertainty. Unfortunately, we do not have the space to discuss this case in detail and will only mention few aspects in Section 4.4. We will provide a more in-depth discussion in Klotz & Shadrova (in preparation).

4.1 Mapping different models into one

We will first discuss the mapping of a number of models into one annotation representation at the example of the realization of aspectual meaning in heritage vs. monolingual speakers of Greek, Russian, and Turkish. The linguistic details are discussed in contribution (Wiese, Allen, et al. 2025 [this volume]). For our purposes, it suffices to understand that aspectual meaning can be encoded in a range of linguistic realizations, which include morphosyntactic (aspectual verb markers), syntactic (periphrastic constructions; subject-drop can play a role), and lexical realizations (verb semantics/Aktionsart; adverbs). These features are diversely distributed between the languages represented in RUEG.

The RUEG perspective predicts systematic shifts in language contact situations, i.e., systematic differences between monolingual speakers and heritage speakers outside of the majority language environment. For example, heritage speakers of different languages might move away from a morphosyntactic realization of aspect and towards a periphrastic or adverbial realization. Of course neither of those categories are identically distributed in the three languages in question in monolinguals and heritage speakers. This means that all aspect realizations in each language need to be annotated. Those are typically well-described in linguistic theory, although the descriptions may not be fully exhaustive due to the open-ended character of language. If there is truly a systematic shift between speaker groups, the data might challenge the existing linguistic models and require adaptations for annotation (see also Section 4.4). Finally, the hypothesis predicts a systematic shift *in a specific direction* for all three languages. To detect this, a comparability between the language-specific descriptions is required. What does it mean, for instance, for Greek aspect to shift from one distribution to another, and (how) can this be quantified in comparison with Russian aspect realization distribution?

Language-specific models of morphosyntactic description will frequently not only differ by the contents of specific categories (e.g. different aspect markers), but also in relevant theoretical assumptions (e.g. UG-based/minimalist grammar vs. usage-based/construction grammar), which may result in drastically different annotations. Will annotation categories for both, or all three, languages hence mean the same, and will the different tag distributions provide conceptual comparability?[12]

The above is only an example of the many mappings required in the annotation of aspect in a corpus like RUEG, as is visualized in Figure 2 in a simplified manner. A comprehensive map of all interrelations would require a reflection of all entities contained by the larger boxes and how they interact with one another. We will not attempt this here. Even in this simplified manner, the modeling requirements for the annotation of a single phenomenon are demanding. This is partially due to the diverse range of language-specific realizations represented in

[12] The fact that linguistic categories are almost never exactly the same between languages has been discussed time and again in typology; see e.g. the controversies and discussion in the construction of the World Atlas of Linguistic Structures (among many others Cysouw et al. 2008) or Croft's (2001) Radical Construction Grammar. There were several attempts to build ontologies for linguistic categories in computational linguistics. The danger here is that the level of abstraction needs to be so high that the tags are no longer informative enough for the problems at hand. The surface-syntactic universal dependencies (SUD; Kahane et al. 2021) annotation we are using for syntax is one case in point – while it may show similarities between languages, it necessarily misses language-specific details we need.

RUEG. However, this multilingual approach only adds another layer of complexity to an already complex map. Even for a single language, a phenomenon that can be realized lexically, morphosyntactically, syntactically, or through other parameters, requires a reflection of all of those linguistic levels separately and in combination; and within the frameworks respectively used. Even if every project member works from the same framework and language, different frameworks and language-specifics may be at work in pre-existing annotations, for instance those created by annotation tools such as syntactic parsers.

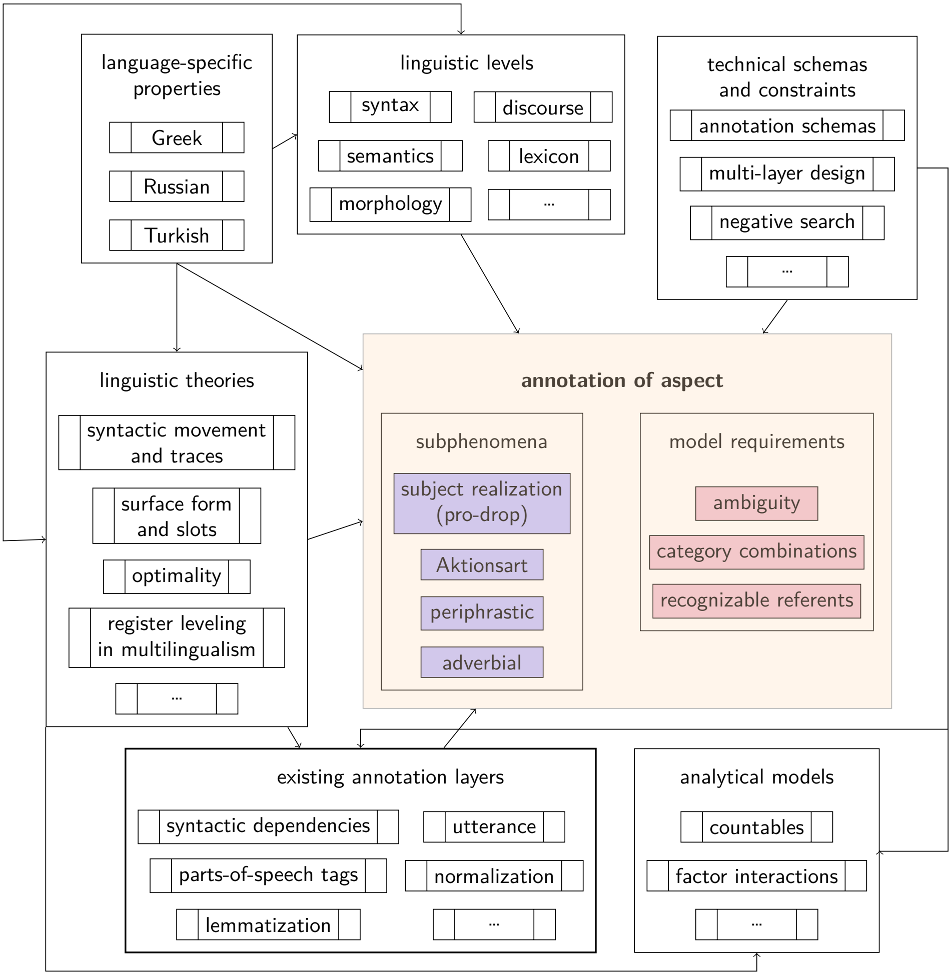

Figure 2: Model interaction in the design of annotations of aspectual expressions in Greek, Russian, and Turkish data

In this way, while the actual annotation aimed for may be definable within the scope of a project (visualized as boxes with singular outlines in Figure 2), most of the models influencing those definitions are not (visualized as boxes with double outlines). As previously mentioned, this has repercussions on the workflow in a corpus project. Expertise is typically distributed by the different fields (technicians vs. linguistic experts) and the subfields they contain (syntax vs. semantics; or Greek vs. Russian). However, a common occurrence in a corpus project like RUEG is that the *specialized* expertise for the data represented in the corpus is acquired primarily by the active annotators. Active annotators may not always have as much experience with the models as designers of the research questions. Neither active annotators nor other linguistic experts often have a full understanding of the computational and technical perspective; and it is not uncommon for neither the active annotators, nor the linguistic experts, nor the technicians to have considered the interface between annotations and the final analytical model envisioned. In the simplest case, this refers to the countable elements and combinations of factors as they need to be made extractable via annotation.

In a best-case scenario, all project members involved would gain awareness of all the interfaces they are handling in the process of devising and performing the relevant annotations. In some sense, annotation can be viewed almost as a separate data collection from existing data.

The result of this work is threefold, as it creates the annotated data, the annotation guidelines, and the procedural and conceptual knowledge generated by all those who participate in annotating and annotation design. The process of annotation thus is a practice of mapping and specifying all of the linguistic models involved to the research question at hand. This requires an ongoing reflection of the perspective the annotated data lends on the existing linguistic models (hence the double-sided arrow in Figure 2). In this lies the unique advantage of corpus linguistics – in their application, we cannot only test linguistic categorizations with respect to their internal consistency, but we also learn on a case-by-case basis what *precisely* our theories have failed to predict. This sound and specific conceptual knowledge around linguistic theories and their combinatorics, i.e., their interfaces, is epistemically rich and beneficial. Conceptualizing annotation and annotation design as actual research, rather than a menial stage of preprocessing, offers many opportunities to clarify the *linguistic* understanding of a phenomenon (see also Shadrova et al. 2025). In spite of this, it is not usually in focus in corpus-linguistic study.

Rather, annotation is often rushed in order to quickly proceed to analysis and interpretation. This is unfortunate for two reasons: One, it puts tremendous pressure on researchers in active annotator roles. In order to annotate well, they have to acquire expertise in all the external models and all the interfaces they are handling, while also acquiring very detailed specialized expertise on the data. They gain rich and useful knowledge that can pave the way to important new research trajectories in the understanding of language; and they acquire critical reasoning skills that are of high value in- and outside of academic employment. In spite of these objective benefits, the work of annotators is often not considered to be "real" research, and the academic recognition is limited, although it is actually of great worth to overall linguistic understanding. Two, when modeling is rushed and interfaces are left undefined, more challenges arise as the lack of clarity carries over to later research stages. This may result in frustrating workflows with many back-and-forth iterations, and can endanger the success of entire projects. This is not to be confused with a workflow that is iterative by design, in which guidelines are developed, tested, and then applied to larger sets. Rather, when interfaces are unintentionally left undefined, annotations are inconsistent, contradictory, and guidelines are changed back and forth, resulting in repeated annotations of the same data *without* increased clarity.

The work of mapping a range of models into a single data space (a corpus) requires an abstraction of the models involved, i.e., a model of each model as is relevant to the current research question. This may imply competing models, which in a multi-layer architecture can be represented in separate annotation layers. However, the outcome of research cannot be fully known in advance, which means that it is impossible to know which details might matter at the end. Thus the abstraction is often avoided by representing as many details of each linguistic theory as possible in the annotation. This, however, only creates more problems by a) burdening annotators with high demands on their working memory; and b) by creating *even more* interfaces between each detail of every model involved, which will then not only need to be reflected in the data (creating bigger data, which can turn out problematic in technical representation especially after reimport into analytical frameworks), but also in the analytical framework itself.

4.2 Mapping partially incompatible models

Mapping multiple models into one to obtain a common linguistic resource poses representational as well as conceptual challenges. Think of a situation (like RUEG's) where an input data model passes through different research groups simultaneously in order to be extended and adapted to the needs of their investigations. The issue of compatibility arises when all output models require unification. A unification procedure generates a combined model from multiple models and has its technical substantiation as a persistent resource, e. g. a linguistic corpus. There is usually more than one unification procedure available. A technical prerequisite for unification is that there is a way of unambiguously representing all models together. In contexts where this is not possible, we deal with representational incompatibility. If it is possible, we may move on to conceptual mappings limiting unification: The output of unification needs to be stable – in the sense that what is represented in a single model is represented again to the same extent and manner in the unified model – and sound, i. e. what emerges from combining each model's aspects together still leads to valid conclusions about the modeled items. If unification fails to be stable and sound, we face conceptual incompatibility.

4.2.1 Representational incompatibility

Failure to represent a unified model is strictly bound to a context, at least for linguistic corpora. For illustration, RUEG's German subcorpus contains two types of syntactic annotations: Syntactic dependencies following the surface-syntactic universal dependencies annotation scheme (SUD; Kahane et al. 2021), and hierarchical topological fields derived from KiDKo's syntax scheme (Wiese et al. 2010), which for the sake of this explanation, can be understood as constituent syntax. While dependency syntax such as SUD appears technically simple, since the input ingredients are tokens with PoS annotations that are enriched with direct and labelled relations between them, the case of constituent syntax seems slightly more complicated: Apart from tokens and PoS annotations, there are non-terminal nodes that do not directly connect to the tokens.[13] The process of annotating dependency syntax is drawing an edge from one token to another and assigning it a label. The process of annotating constituent syntax, though, includes defining non-terminal nodes and the adjecent member tokens or non-terminal node members of it. Thus, the demands towards an annotation tool are quite distinct, and tools which can do both are quite rare. The representation of those two types of annotations and thus an underlying unified model of both constituent and dependency syntax on one screen fails, i. e. the models are incompatible in the context of the actual creation process. Similarly, it holds for many linguistic data formats, whose representational capabilities are usually a reflection of the features of an annotation tool, that they can either represent dependency syntax or constituent syntax, but not both together. Some (but not the only) exceptions are the TCF format of the online pipelining tool Weblicht

[13] Alternatively, the scheme introduces empty nodes.

(Hinrichs et al. 2010), serializations of UIMA CAS (Ogren & Bethard 2009), or the annotation tool Hexatomic (Druskat et al. 2023). In the context of search and visualization – or information retrieval – combining a model of dependency syntax and constituent syntax succeeds through apt representations (Krause & Zeldes 2014) and unification procedures (Zipser & Romary 2010, Fei et al. 2021, Krause & Klotz 2023).

4.2.2 Conceptual incompatibility

Some problems that appear to be technical issues on the surface reveal themselves to be more fundamental conceptual incompatibilities at closer inspection. It simply takes a rigid formal environment for underlying conceptual issues, that might otherwise have never been detected, to come to light. To understand under which conditions models are conceptually incompatible, we need to look at the circumstances under which their unification might be unstable or unsound. In our example of two syntactic schemes, we can think of a unification procedure that leads to different results about our annotated language samples in a unified model compared to the dependency model only. Unifying dependency and constituent trees in one model must answer the question on how to proceed with non-terminal nodes when defining the common set of tokens. Non-terminal nodes in constituent syntax can be understood as empty nodes, since they do not connect to an actual token. A unification procedure that does not map non-terminals directly to tokens, but models them as indirectly connected through their descendent nodes, is stable. It does not affect dependency trees and what can be revealed about them.[14] An alternative unification procedure, though, that maps non-terminal constituent nodes to newly created token objects with an empty text value between the existing input token objects, does. Even though we can still map the original dependency trees somehow onto all previously existing tokens, since none of them was deleted: By inserting new token objects, the unification alters what will be revealed about the linear order of tokens, e. g. in an n-gram analysis. Thus it is not stable.

Incompatibility arising from unsound unification is the more dangerous one, since it easily goes unnoticed. Representational compatibility is already given and we might not realize that one of our unification's mappings produces an inaccurate model of our data. Parts of the Russian and English spoken RUEG data have been annotated for prosodic features; pitch accent and boundary tones (Zerbian et al. 2022, 2024). The same subset of the data has simultaneously been an-

[14] It furthermore is sound, since everything we learn about dependence structures in combination with constituents is true.

notated for basic functional and lexical features (e. g. PoS and lemma, henceforth subsumed as *morphology* or *morphological*) by another group. The shared input model to both remodeling processes, morphology and prosody, is a sequence of sub-segmentations and segmentation into tokens of an audio stream. While the morphology group bases their annotations on normalized tokens, which are a reinterpretation of the input tokens, the prosodic annotations require a subdivision of the input tokens into syllables. Normalization includes orthographic correction, split, and merge of tokens. Subdivision into syllables splits an input token at least zero times (one-syllable tokens). Both output models still contain the shared input model's entities and add new base units and annotations on top. They are illustrated on the left in Figure 3 along the example *everyone's*, a merged form of *everyone* with a clitic form of *is*, which in the diplomatic input is a single token. The prosodic group divides it into syllables ɛ, vɹi, and wənz and assigns a pitch accent to the first. For morphology, the input token is normalized as two tokens *everyone* and *is*, of which the former is annotated to be a pronoun, the latter to be an auxiliary. Associations of annotations with their annotated unit is visualized as horizontal overlap. At first glance, the two models' representations seem comparable and thus compatible. An attempted unification is illustrated in Figure 3 on the right. We can easily diagnose stability, the content of each model completely transfers to the unified model. Unfortunately, the unified model is not sound. By the models logic of association by overlap, we actually do not know how to map syllables onto normalized tokens, but the model requires us to provide such a mapping for unification. How annotations for normalized tokens map onto syllables is undefined, so by guessing we end up with wrong prosodic information on lexical items and vice versa.

pos	PRON	AUX
norm	everyone	is
dipl	everyone's	

dipl	everyone's		
syl	ɛ	vɹi	wənz
pitch-acc	+		

pos	PRON			AUX
norm	everyone			is
dipl	everyone's			
syl	ɛ	vɹi	wənz	
pitch-acc	+			

Figure 3: Two input models (on the left) are unified in a stable, but unsound manner (on the right). The mapping between syllables (`syl`) and normed (`norm`) segments is undefined. The example is taken from speaker and situation `USbi53MR_isE`.

Luckily, the conceptual incompatibility of the prosodic and morphological model can be resolved by updating the unification procedure to the point that it does not require a mapping of syllables to normalized tokens. A mapping is initially required, because the logic of association by overlap is transitive.[15] But there is no way to map the two annotation target groups `syl` and `norm` onto each other in a sound way, that keeps our unification stable, as long as we rely on a transitive relation between entities in our model and as long as we do not add a common minimal segmentation of both, `norm` and `syl`. The solution in Figure 4 provides a minimal segmentation layer (`min`)[16], to which syllables and normalized tokens refer, thus a mapping between the two is given. In practice such a minimal layer is expensive in terms of annotation effort, and not necessarily defined for all potential cases. However, if we replace the transitive association relation by a semantically weaker alignment pointer, a more feasible solution can be achieved (cf. Figure 5). This way, both input models keep a copy of their diplomatic tokens and we align those input segments as if they were a parallel corpus. This way, we avoid transferring morphological features to syllables. Also, overlap is not given anymore and cannot transitively transfer morphological features to prosodic units. As a consequence, we do not require a mapping between them.

pos	PRON								AUX	
norm	everyone								is	
dipl	everyone’s									
min	e	v	e	r	y	o	n	e	’	s
syl	ɛ	vɹi				wənz				
pitch-acc	+									

Figure 4: A stable and sound unification: A common minimal basic tokenization layer for morphology and prosody leads to a clean mapping between all annotations. Nevertheless, such a mapping is expensive, disputable, and not always available.

Conceptual incompatibility must not be mistaken for theoretical contradiction. A not very faithful syntactician might very well use dependency and constituent syntax together when annotating their data, because they see the two schemes’ advantages adding up very well in their favour; so said researcher unifies two

[15] Here we refer to transitivity in the mathematical sense: A relation that applies for two entities *a* and *b* as well as *b* and *c* thus also applies to *a* and *c*.

[16] The layer is minimal w. r. t. to layers `syl` and `norm` since each unit in one of the two latter layers can be modelled as a sequence of one or more units of said layer.

pos	PRON	AUX
norm	everyone	is
dipl	everyone’s	

↑

dipl	everyone’s		
syl	ɛ	vɹi	wənz
pitch-acc	+		

Figure 5: A sound and stable unification: A parallel corpus architecture maps only the common base units onto each other and keeps prosodic and morphological layers aligned, but separated.

data models of different syntactic annotations. Annotations are first and foremost a means of retrieving information and must not be dismissed for not being what they were never meant to be (a linguistic theory). Rather, they need to be understood as a (unified) model and in this an interpretation and reduction. At the same time, we cannot take this view on annotations so far as to consider them labels without any theoretical underpinning (Bubenhofer 2018). Especially in the context of multi-layer architecture, the theoretical underpinnings gain importance from the process perspective of corpus creation (meaning all of the processes that lead to one released corpus) being a prerequisite of successful information retrieval. This is again exemplified in context of the RUEG project in Section 4.3 below.

4.3 Interdependence between annotation layers

Designing annotation layers, especially lower-level layers containing basic information such as normalization, lemmatization, or PoS, always constrains which dependent annotation layers can be defined on their basis later on. Usually, in corpus linguistics, we are then interested in higher-level concepts that use these layers for their domain of annotation, i. e. they directly or indirectly depend on them. This dependence implies that designing and annotating lower-level layers requires transparent annotation decisions and a diligent annotation procedure. At the same time, the dependence relations between higher-level layers and lower-level layers always need to be explicitly and formally documented.

The terms *dependent (annotation) layer* and *independent (annotation) layer* will be used in the following. The *dependent layer* denotes the layer that is based on a layer added to the corpus earlier. The *independent layer* is that earlier annotated layer another layer builds on. This does not mean the independent layer cannot itself be dependent on any of the other layers in the corpus. It only makes a statement about relations between two layers within one pairwise comparison. Since annotations necessarily rely on specific underlying models, interdependence between annotation layers is not only relevant from a technical perspective. Ignoring interdependence between annotation layers can lead to at least two different problems:

Problem A: Annotation layers are treated as independent of each other while they are not.

Problem B: Inaccurate annotations on one annotation layer cause inaccuracies on a dependent annotation layer created and annotated later on.

Both problems are strongly intertwined in the sense that if problem A is not avoided, problem B will occur as well. Hence, both are illustrated using the same example of syntax annotations of the RUEG data. Such annotations exist for all RUEG languages. Syntactic trees following the Universal Dependencies scheme for syntactic annotations (UD; de Marneffe et al. 2021) are available for Turkish, English and Greek. These annotations did not undergo a step of manual correction. For Russian and German, there are manually corrected syntactic annotations following the SUD annotation scheme. The sequence of successive processes that led to the different syntax annotations is visualized in Figure 6. It shows only the section of layers and dependences between them that are relevant here.

Complex interdependence between annotation layers already unfolds in the process product of SUD annotations alone. A wrong syntactic relation or wrong relation label in the automatically preparsed SUD annotations could originate in any of the layers the SUD layer directly or indirectly depends on. An annotator correcting such parses needs to be aware of the existence and the nature of the interdependent structure to know where to look for possible errors leading to an error on the layer that is of current concern.

Newly emerging interdependences are one reason why annotation should be viewed as an iterative process. Including a new annotation layer in the corpus that is based on a now emerged independent layer may cause annotation decisions concerning the independent layer to be reconsidered and annotations to be revised and adjusted accordingly. For instance, during correction of automatically parsed syntactic annotations, an annotator using a top-down perspective may detect a mismatch between the assigned PoS tags of two tokens and the requirements of the syntactic relation they would like to assign for them. What

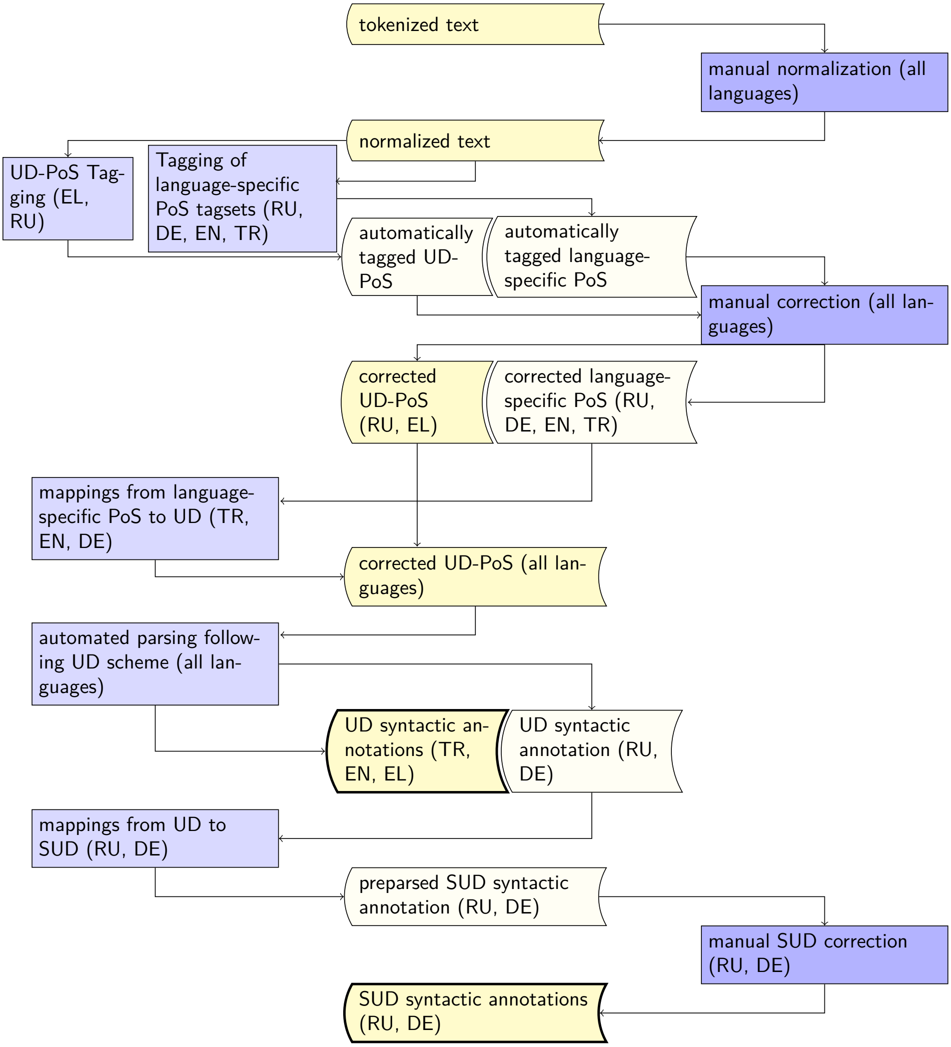
tokenized text
manual normalization (all languages)
normalized text
UD-PoS Tagging (EL, RU)
Tagging of language-specific PoS tagsets (RU, DE, EN, TR)
automatically tagged UD-PoS
automatically tagged language-specific PoS
manual correction (all languages)
corrected UD-PoS (RU, EL)
corrected language-specific PoS (RU, DE, EN, TR)
mappings from language-specific PoS to UD (TR, EN, DE)
corrected UD-PoS (all languages)
automated parsing following UD scheme (all languages)
UD syntactic annotations (TR, EN, EL)
UD syntactic annotation (RU, DE)
mappings from UD to SUD (RU, DE)
preparsed SUD syntactic annotation (RU, DE)
manual SUD correction (RU, DE)
SUD syntactic annotations (RU, DE)

Figure 6: Subprocesses involved in the creation of syntactic dependency annotations for all RUEG languages. Book shaped boxes indicate process products (light yellow – intermediate process product; brighter yellow – process product.) Rectangle shaped boxes represent processes (light violet – automated process; brighter violet – manual process). Arrows indicate *results in* or *is passed over to*.

could be a simple annotation error on the PoS layer might also indicate a general mismatch between the models operationalized in the PoS tagset (or lower-level layers) on the one hand and in the label set for syntactic dependencies on the other hand. The PoS annotation might require revision, the definition of the syntactic relations might need to be widened to describe the data, or there may be unresolved conflicts with other annotation layers, the syntactic annotation indirectly relies on, e. g. a tokenization process that already creates preconditions for applying the "correct" dependence relation.

As formulated in Problem B, inaccurate annotations on the independent layer get inherited by a dependent layer in automatic pre-annotations, which increases the effort of manual post-correction. Otherwise the result is an annotation layer with numerous incorrect annotation values. Unawareness of an interdependence between syntactic relations and PoS however, i. e. problem A, would mean that an annotator does not take the PoS layer into account during the correction of syntactic annotations. Potential mismatches between models are overlooked and necessary model adaptions may be missed. This, again, would result in either incomprehensible or misinterpreted outcomes in analysis. This can lead to false interpretations and assumptions for our linguistic understanding, particularly when looking for emerging phenomena. A lack of awareness of attention towards the interdependence of annotation layers results in an underestimation of the complexity of the applied model(s) and of the problem at hand. Future annotations, that then rely on such annotations, fail to be understood from the beginning. This way, an initial misconception is passed on from annotation to annotation and is mirrored in any follow-up analysis. For instance, if we fail to identify proper nouns in PoS, or ignore PoS in our annotations of syntax, we may then also fail to evaluate the introduction of referents through specific syntactic means.

Ignoring the interdependence of layers replicates itself when more additional annotation layers are added while previous interdependence still goes unnoticed, undocumented and not taken into consideration for the planning of new dependent layers. An alternative approach to documenting interdependence and the consequences arising from that for the underlying models is possible if different models cannot be conceptually reconciled. In this case the independence and incompatibility of the layers has to be made transparent and explicit in the documentation because incompatible layers should consequently not be used together for analyses. A more radical solution is to publish such layers separately in different subcorpora.

4.4 Mapping unknown models

So far, we have discussed different aspects of handling multiple models, such as their compatibility or their interdependence. A special case occurs in the modeling of data from fully or partially *unknown* linguistic models. As RUEG is interested in *emergent* properties, i.e., new linguistic feature distributions that cannot be fully predicted from the sum of the parts of other features, this is a common requirement in the project context.

The linguistic basis for the process we sketch out in this section is the idea that language behaves as a complex dynamic, i.e. emergent, system (Beckner et al. 2009, Ellis & Larsen-Freeman 2009, Massip-Bonet 2013, Croft 2014: among others). A system of this kind is emergent, because it amounts to more than the sums of its parts and their relations, i.e. some effects arise from the interaction of factors or even interactions. It is (mathematically) *complex*, because the behavior of the entire system can be stochastically described, but not fully predicted. It is dynamic, because it moves through different states in an ever-changing and evolving way. Dynamics may include *directional* movement (loss or emergence of cases) or oscillating movement (loss and re-emergence of negation markers), and processes of grammaticalization and de-grammaticalization equally. Language has been modeled as a complex dynamic system both on population (German in Germany) and individual level (a speaker's linguistic system at various stages of acquisition or years of life, see Hiver et al. (2022) for a recent overview).

Some of the underlying assumptions of the dynamic systems/emergentist perspective on language change and language development are that although shifts happen systematically, they are not fully predictable; that once a relevant shift happens to the system, other paths are less likely to be taken (path dependence); that shifts do not usually encompass the entire system at once, but can have *repercussions* on the entire system (dialectics); that shifts in minor aspects of a language may cascade into more widely effective changes (quantity-quality shift); and that amidst a changing process, the system will be highly variable in the relevant aspects due to residing in transitory, superpositioned, and partially unpredictable states. These may later resolve in one direction or another; however, in describing a system from within its transitory state, different possibilities have to be considered at the same time.

These transitory states are what we refer to as partially (or fully) unknown models. Although they might eventually be fully describable (i.e. known), during analysis they require a perspective that is not fully guided by the origin or the presumed target of the development. Rather, we must assume that the abstractions we are looking for would differ from the canonical paradigm of the standard language, but since it is new, the details are unknown.

We might try to solve this by defining target hypotheses (Reznicek et al. 2013, Lüdeling & Hirschmann 2015, Laarmann-Quante et al. 2017), i.e., a normalized layer of the utterance in question, as it is common in learner corpus research. For this, one would note the canonical form next to the form evidenced in the corpus and classify the differences. However, this does not provide a solution for the emergent system, as the target is not the canonical form – rather, this approach would unintentionally result in a deficiency-oriented perspective, in which one presumes that heritage speakers "mean", i.e., project as their target, the canonical standard language, but cannot fully meet its requirements. If one were to analyze the new paradigm based on target hypotheses, one would inevitably analyze the emergent system as merely a divergence from the state it emerged from, projecting the old logic and inner constraints onto the system.

Alternatively, one could try and analyze the new forms based on *an assumption* of what the new system might look like. This assumption may be highly educated, for instance informed by theories of ongoing language change and the dynamics of bilingualism. However, even an educated guess is *only a first hypothesis*, not a synthesized and integrated theory of the new paradigm. It will likely not match the data in all relevant aspects. If the presumed new paradigm has already been projected onto the data in the form of annotations, one enters a risk of getting trapped in a never-ending cycle of re-annotating: as soon as one finds a counter-example, the entire new system has to be reworked, hence requiring re-annotation of all previous data. However, since one's initial hypothesis was probably right in some aspects, then, evidence may be gathered again for the initial hypothesis. The hypothesized system bounces back and the annotations have to be reworked again. It is important to note that this is different from an iterative annotation process based on a clearly defined linguistic model, because the limits and degree of certainty of the hypothesized model (the educated guess) are unknown at each step. It follows that the next iteration does not necessarily result in greater certainty or clarity, as is the case in improved annotation guidelines. Rather, the annotation guidelines and the target system oscillate between several equally uncertain systems.

A solution to this problem instead lies in modeling by form (in a corpus-driven manner) when abstractions are not yet sufficiently certain to be included in the annotation, and mapping relevant aspects of the *original*, i.e., canonical forms, in a separate layer. In this new way, the mapping represents the observed data *form* mapped to a presumed target *function*, i.e., the mapping between the new form and the old function is preserved, but no new function is presumed. Unlike this, the *functional* annotation of the presumed function is *more abstract* than in the approaches mentioned above.

Careful consideration of the effect of abstraction at each step is required in order to arrive at precise annotations. This is particularly the case where models themselves are unstable or dynamic. Finding the right timing and the right level of abstraction requires a bird's eye view on all processes and models involved and is usually not achievable ad hoc, but itself an iterative and communicatively demanding task. However, it is the only way to derive an emergent phenomenon from existing data, and, if done carefully, provides deep insights into all aspects of the target paradigm including its dynamic and not yet fully shifted parts. With this, we conclude our discussion of different types of model and interface mappings.

5 Discussion

Corpus data is rather unique in linguistic research, as it contains language at full complexity regardless of how narrowly defined the communicative situation of its elicitation. Unlike in an experimental setting, complexity is not reduced in a selection process prior to elicitation, and outcomes are not funneled into one of *n* predetermined options. This is why all the modeling and analytical design has to be completed after elicitation. An aspect that we cannot discuss in this contribution, but that also concerns similar modeling decisions, is that of mapping quantitative models to the open-ended data that corpora present. These mappings, too, reflect complexes of explicit or implicit analytical decisions that carry repercussions for the evidential scope[17] of corpus-linguistic research.

Most of the epistemological weight of corpus-linguistic research lies on post hoc differentiation, rather than a priori design, of the data – it is during annotation design and execution, data extraction (i.e., selection), and data analysis that the fitting of all involved linguistic models with the actual data happens. In addition to the linguistic complexities, a range of technical limitations poses further demands.

In an ideal setting, all of the involved models are defined somewhere (in a linguistic theory outline; a technical framework manual; or an annotation guideline). Unpacking this complexity requires manual, additive, and collaborative work, and effectively creates distributed knowledge. This means that even in the best of scenarios, there are no natural carriers of *all* the information regarding *all* the models relevant to a strongly collaborative project, or simply put: no-one knows everything that matters about the data. This is commonly solved by

[17] The evidential scope (see e. g. Leonelli 2009, 2019) defines the limits of what can be learned from a specific type of data or methodology.

distributing work between representatives of different fields of expertise, for instance by research field, methodology, or support roles (technicians, statisticians, linguists, etc.).

Although it lifts some of the pressure to understand many models, this division of labor creates further interfaces between those who carry the understanding of some aspects and those who carry the understanding of some other aspects. In that context, setting aside space for a duplicate of each set of epistemes (i.e., two people working on the same model subset) would be helpful, but is not highly realistic under the current conditions of research funding.

Worse perhaps, no-one can know *enough* about the data a priori in a corpus project, since it is unpredictably variable even in the narrowest of settings. Expertise thus has to be grown *from within the corpus project*, that is from a growing understanding of the data, the specific requirements of the annotation or analytical model, and the technological options and limitations. These epistemes emerge as a process product of corpus-linguistic research. Even with high discipline and the best of intentions, they can be difficult to document and transfer, because much of the procedural knowledge exists on a perceptual level which can be difficult to notice and describe, and thus it often remains implicit. Additionally, keeping a discipline at documenting can be a challenge due to the high number of interfaces and mental objects involved, and since there is very little incentive within the academic system for it. Luckily, this is slowly beginning to change. In recent years, well-designed and documented resources and research software are more commonly acknowledged as academic output. This is evidenced, for instance, by the increase in published annotation guidelines, the option to publish materials and documentations as data supplements with many research articles, as well as a changed CV structure in some funding agencies and the professionalization of research infrastructure citation (Lavoie 2012, Anzt et al. 2020, Schlauch et al. 2022, among others). In spite of these very positive developments, the fact remains that the kind of procedural knowledge we have described is not yet regarded in the same way as more established formats of scientific publishing, although it is often the first and primary type of knowledge gained in a complex corpus project during compilation (i.e., before final analysis).

While it seems clear that procedural knowledge relevant to a particular project cannot be simply copied onto the next set of problems, the transfer of process-oriented knowledge is crucial to the success of each corpus project, which includes *every collaborator's next project*. Personal continuity is helpful, but it is clear that not every single person working on a corpus project will continue on the same path in the future; and not every new collaborator in a future corpus project would prefer to read through the documentations of every single other, even similar, corpus project. Even if they were disciplined and motivated to do so, the number of mental objects, interface definitions and so on would easily occur as overwhelming to them as they would have been to the initial group.

(How) can the *understanding* of model fitting then be transferred between projects? Would it require more synthesis of existing procedural knowledge, as we have attempted in this contribution – a theoretical corpus linguistics for a phenomenology and a typology of corpus-specific modeling problems? Or does it suffice to know the linguistic details and the general necessity to make them fit with the quantitative and technical models and frameworks? Should corpus-linguistic infrastructure and modeling be viewed as a type of advisory work, similar to a statistical consultant, or as a type of engineering work akin to research software engineering; or would *all* corpus linguists need to be more involved with the quantitative, technical and also epistemological aspects of their work, resulting in a community-of-practice approach, in which all participants understand the majority of their data in all its aspects and develop the expertise to debate it in multidimensional ways? If so, as a logical consequence their work would require smaller corpora with more time and space for in-depth linguistic and modeling debate.

With its focus on deep annotation, clear theoretical embedding, an integration and comparison of various features, and its strict handling of documentation, quality assurance, and continuous integration, RUEG is a good example for best practices in some aspects of this. It has created an impressively well-documented, impressively well-controlled and rich database which will not only continue to be used in future research in its current state, but can also be enriched with new layers of annotation, expanded, and its materials reused for conceptual replication. However, even at this level of precision and disciplined documentation, there are still manifold interfaces that will in all likelihood remain underspecified and untreated, unless they are attended to in other projects; and it is not highly likely to receive funding for the modeling of legacy data from other projects. This suggests that RUEG likely poses an upper limit to what can be contained within a corpus project without creating insurmountable organizational and communicative overhead.

This tension stands a huge challenge for prospective projects, especially within the context of a strong emphasis of a multifactorial quantitative paradigm in the current debate. Corpus complexity is a direct correlate of the number of models and factor combinations involved: the more factors there are to be controlled for, the more factors will need to be considered in the analytical model. This results in more variability, in this case variability along more dimensions, that need to be accounted for in the data. To linguistically capture the full extent of the data, a more complex model is required, i.e., the inclusion of more linguistic aspects or sub-models and their interactions. All of these aspects – the factors, the variability along each dimension, each model, the hypotheses concerning with respect to every aspect, the multidimensional results and their repercussions for each of the involved linguistic subdomains – then has to be considered and discussed in the scientific debate, if as many epistemes as possible are to be procured in the process. From the complexity emerging from this interaction of elements arises a big epistemological question: what can reasonably be *understood* and *explained* as *linguistic knowledge*, rather than simply quantitatively *described* in naturalistic data?

Although task-based corpora have been used for many years in some subfields of corpus linguistics, a recent shift has taken the paradigm towards the elicitation of data that is very strictly controlled by a range of factors. This is a reflection of a better understanding of how task and register effects shape natural language and is a generally positive development, as these elicitations have lead to a much better understanding of the multiplicity of interactions of factors and factor combinations in language production (e.g. Alexopoulou et al. 2017, see also Wiese, Labrenz, et al. 2025 [this volume]). This newer tendency, of which RUEG is one example, has also shown that speakers in *any* communicative situation do not produce (quasi-)experimental data. Rather, we see that the full scale of emergent properties of natural language production and the multiplicity of interrelations between linguistic and extra-linguistic factors are present even within such narrowly constrained communicative situations. It is only natural to conclude that corpus linguistics should pay close attention to those emergent properties. With this contribution, we argue for more exchange, debate, and synthesis of corpus-based modeling expertise, as the many interwoven layers of the emergent space require much epistemological clarity to be fully grasped and productively handled.

Abbreviations

PoS	part of speech
SUD	surface-syntactic universal dependencies
UD	universal dependencies

Acknowledgements

All authors are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Research Unit *Emerging Grammars*, Projects Pd *A* *cross-linguistic corpus of comparative data in heritage and majority language use*, 394844736, and Pc *Corpus-linguistic methods*, 313607803.

We thank Thomas Krause for his crucial role in developing the conceptual and practical foundations of the corpus infrastructure. We are grateful to our student assistant Esra Uyanık for her support in corpus development and careful copy-editing of this chapter. We express our gratitude to Sabine Zerbian, Maria Martynova, Mareike Keller, Nadine Zürn, and two anonymous reviewers for their insightful feedback which has helped to improve this chapter greatly. We also thank Lea Coy for help with pre-publication formatting.

None of this would have been possible without the inspiring collaboration with and the support of the entire RUEG team and the Mercator Fellows.

Additionally, Anke Lüdeling and Martin Klotz acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412 *Register*, 416591334.

References

Alexopoulou, Theodora, Marije Michel, Akira Murakami & Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1). 180–208. DOI: 10.1111/lang.12232.

Anzt, Hartwig, Felix Bach, Stephan Druskat, Frank Löffler, Axel Loewe, Bernhard Y Renard, Gunnar Seemann, Alexander Struck, Elke Achhammer, Piush Aggarwal, et al. 2020. An environment for sustainable research software in Germany and beyond: Current state, open challenges, and call for action. *F1000Research* 9. DOI: 10.12688/f1000research.23224.2.

Atwell, Eric. 2008. Development of tagsets for part-of-speech tagging. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 501–527. Berlin: De Gruyter Mouton.

Bauer, Laurie. 2000. Word. In Geert E. Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphologie/Morphology: Ein internationales Handbuch zur Flexion und Wortbildung/An international handbook on inflection and word formation*, vol. 17, 247–257. De Gruyter Mouton. DOI: 10.1515/9783110111286.1.4.247.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, et al. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59. 1–26. DOI: 10.1111/j.1467-9922.2009.00533.x.

Biber, Douglas. 2015. Corpus-based and corpus-driven analyses of language variation and use. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 193–224. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199544004.013.0008.

Boersma, Paul. 2001. Praat: A system for doing phonetics by computer. *Glot International* 5(9/10). 341–345.

Bubenhofer, Noah. 2018. Wenn „Linguistik“ in „Korpuslinguistik“ bedeutungslos wird: Vier Thesen zur Zukunft der Korpuslinguistik. *Osnabrücker Beiträge zur Sprachtheorie (OBST)* (32). 17–30. DOI: 10.5167/uzh-199167.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780198299554.001.0001.

Croft, William. 2014. Studying language as a complex adaptive system. *English Linguistics* 31(1). 1–21. DOI: 10.9793/elsj.31.1_1.

Cysouw, Michael, Mihai Albu & Andreas Dress. 2008. Analyzing feature consistency using dissimilarity matrices. *Sprachtypologie und Universalienforschung* 61(3). 263–279. DOI: 10.1524/stuf.2008.0025.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Dewaele, Jean-Marc. 2019. The vital need for ontological, epistemological and methodological diversity in applied linguistics. In Claire Wright, Lou Harvey & James Simpson (eds.), *Voices and practices in applied linguistics: Diversifying a discipline*, 71–88. York: White Rose University Press. DOI: 10.22599/BAAL1.e.

Druskat, Stephan, Thomas Krause, Clara Lachenmaier & Bastian Bunzeck. 2023. Hexatomic: An extensible, OS-independent platform for deep multi-layer linguistic annotation of corpora. *Journal of Open Source Software* 8(86). 4825. DOI: 10.21105/joss.04825.

Ellis, Nick C. & Diane Larsen-Freeman. 2009. *Language as a complex adaptive system*. John Wiley & Sons.

Fei, Hao, Shengqiong Wu, Yafeng Ren, Fei Li & Donghong Ji. 2021. Better combine them together! Integrating syntactic constituency and dependency representations for semantic role labeling. In Chengqing Zong, Fei Xia, Wenjie Li & Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 549–559. Stroudsburg, PA: Association for Computational Linguistics. DOI: 10.18653/v1/2021.findings-acl.49.

Gould, Elliot, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, et al. 2025. Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology* 23(35). DOI: 10.1186/s12915-024-02101-x.

Gozli, Davood. 2019. Shifting focus. *Experimental Psychology and Human Agency*. 1–16. DOI: 10.1007/978-3-030-20422-8_1.

Hinrichs, Erhard W., Marie Hinrichs & Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In Sandra Kübler (ed.), *Proceedings of the ACL 2010 system demonstrations*, 25–29. Uppsala: Association for Computational Linguistics. http://www.aclweb.org/anthology/P10-4005.

Hiver, Phil, Ali H Al-Hoorie & Reid Evans. 2022. Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition* 44. 913–141. DOI: 10.1017/S0272263121000553.

Kahane, Sylvain, Bernard Caron, Emmett Strickland & Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In Daniel Dakota, Kilian Evang & Sandra Kübler (eds.), *Proceedings of the 20th international workshop on treebanks and linguistic theories (TLT, SyntaxFest 2021)*, 35–47. Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/2021.tlt-1.4 (24 June, 2025).

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

Klotz, Martin & Anna Shadrova. In preparation. *The impact of manual post correction of annotations on corpus analysis*.

Knobloch, Clemens & Burkhard Schaeder. 2000. Kriterien für die Definition von Wortarten. In Geert E. Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphologie/Morphology: Ein internationales Handbuch zur Flexion und Wortbildung/An international handbook on inflection and word formation*, vol. 1, 674–692. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110111286.1.10.674.

Krause, Thomas & Martin Klotz. 2023. *Annatto*. Version 0.2.0. https://github.com/korpling/annatto/.

Krause, Thomas & Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1). 118–139. DOI: 10.1093/llc/fqu057.

Laarmann-Quante, Ronja, Katrin Ortmann, Anna Ehlert, Maurice Vogel & Stefanie Dipper. 2017. Annotating orthographic target hypotheses in a German L1 learner corpus. In Joel Tetreault, Jill Burstein, Claudia Leacock & Helen Yannakoudakis (eds.), *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, 444–456. Copenhagen: Association for Computational Linguistics. DOI: 10.18653/v1/W17-5051.

Lavoie, Brian F. 2012. Sustainable research data. *Managing Research Data.* 67–82. DOI: 10.29085/9781856048910.005.

Leonelli, Sabina. 2009. On the locality of data and claims about phenomena. *Philosophy of Science* 76(5). 737–749. DOI: 10.1086/605804.

Leonelli, Sabina. 2019. What distinguishes data from models? *European Journal for Philosophy of Science* 9(2). 22. DOI: 10.1007/s13194-018-0246-0.

Lüdeling, Anke, Artemis Alexiadou, Aria Adli, Karin Donhauser, Malte Dreyer, Markus Egg, Anna Helene Feulner, Natalia Gagarina, Wolfgang Hock, Stefanie Jannedy, Frank Kammerzell, Pia Knoeferle, Thomas Krause, Manfred Krifka, Silvia Kutscher, Beate Lütke, Thomas McFadden, Roland Meyer, Christine Mooshammer, Stefan Müller, Katja Maquate, Muriel Norde, Uli Sauerland, Stephanie Solt, Luka Szucsich, Elisabeth Verhoeven, Richard Waltereit, Anne Wolfsgruber & Lars Erik Zeige. 2022. Register: Language users' knowledge of situational-functional variation. *Register Aspects of Language in Situation* 1(1). 1–58. DOI: 10.18452/24901.

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

Lüdeling, Anke & Hagen Hirschmann. 2015. Error annotation systems. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge handbook of learner corpus research*, 135–157. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139649414.007.

Massip-Bonet, Àngels. 2013. Language as a complex adaptive system: Towards an integrative linguistics. In Àngels Massip-Bonet & Albert Bastardas-Boada (eds.), *Complexity perspectives on language, communication and society*, 35–60. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-32817-6_4.

Moisl, Hermann. 2009. Exploratory multivariate analysis. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 874–899. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.874.

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

Ogren, Philip V. & Steven J. Bethard. 2009. Building test suites for UIMA components. In Kevin Bretonnel Cohen & Marc Light (eds.), *Proceedings of the workshop on software engineering, testing, and quality assurance for Natural Lan-*

guage Processing (SETQA-NLP 2009), 1–4. Boulder, CO.: Association for Computational Linguistics. https://aclanthology.org/W09-1501/.

Özsoy, Onur & Frederic Blum. 2023. Exploring individual variation in Turkish heritage speakers' complex linguistic productions: Evidence from discourse markers. *Applied Psycholinguistics* 44(4). 534–564. DOI: 10.1017/S0142716423000267.

Pascual y Cabo, Diego & Jason Rothman. 2012. The (il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Applied Linguistics* 33(4). 450–455. DOI: 10.1093/applin/ams037.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eight international conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

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

Reznicek, Marc, Anke Lüdeling & Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus. *Automatic Treatment and Analysis of Learner Corpus Data* 59. 101–123. DOI: 10.1075/scl.59.07rez.

Robins, Robert H. 1986. The Technê Grammatikê of Dionysius Thrax in its historical perspective: The evolution of the traditional European word class system. In Pierre Swiggers & Willy van Hoecke (eds.), *Mot et parties du discours. Word and word classes. Wort und Wortarten*. Leuven: Leuven University Press.

Rothman, Jason. 2009. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism* 13(2). 155–163. DOI: 10.1177/1367006909339814.

Schlauch, Tobias, Oliver Bertuch, Oliver Knodel, Guido Juckeland, Stephan Druskat & Michael Meinel. 2022. HERMES: Automated software publication with rich metadata. In *Helmholtz open science briefing: Helmholtz open science forum Forschungssoftware*, 107–116. DOI: 10.5281/zenodo.6241553.

Schmid, Helmut. 2008. Tokenizing and part-of-speech tagging. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, vol. 1 (HSK), chap. 24, 527–551. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110211429.1.484.

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 402–419. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199571932.013.030.

Shadrova, Anna. 2025. No three productions alike: Lexical variability, situated dynamics, and path dependence in task-based corpora. *Open Linguistics* 11(1). 20240036. DOI: 10.1515/opli-2024-0036.

Shadrova, Anna, Joshua Boivin, Rahel Gajaneh Hartz & Anke Lüdeling. In preparation. *Bilingualism modulates complex verb use in a task-based corpus of L1 German.*

Shadrova, Anna, Anke Lüdeling, Rahel Gajaneh Hartz, Thomas Krause & Martin Klotz. 2025. „Step away from the computer!“: Über die linguistische Datenkategorisierung als Erkenntnisprozess und daraus folgende Herausforderungen bei der Nachnutzung von Annotationen und Annotationstools. *Zeitschrift für Germanistische Linguistik* 53(1). 166–214.

Stachowiak, Herbert. 1973. *Allgemeine Modelltheorie.* Wien & New York: Springer.

White, Hayden. 1973. *Metahistory: The historical imagination in nineteenth-century Europe.* John Hopkins University.

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

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

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

Wiese, Heike, Oliver Bunk, Fynn Dobler, Ulrike Freywald, Sophie Hamm, Banu Hueck, Anne Junghans, Jana Kiolbassa, Julia Kostka, Marlen Leisner, Nadine Lestmann, Katharina Mayr, Tiner Özçelik, Charlotte Pauli, Gergana Popova, Ines Rehbein, Nadja Reinhold, Franziska Rohland, Sören Schalowski, Kathleen Schumann, Kristina T. Sommer & Emiel Visser. 2010. *KiDKo: Ein Korpus spontaner Unterhaltungen unter Jugendlichen im multiethnischen und monoethnischen urbanen Raum.* http://kiezdeutschkorpus.de/en/.

Wiese, Heike, Annika Labrenz & Albrun Roy. 2025. Tapping into speakers’ repertoires: Elicitation of register-differentiated productions across groups. In Shanley E. M. Allen, Mareike Keller, Artemis Alexiadou & Heike Wiese (eds.), *Lin-*

guistic dynamics in heritage speakers: Insights from the RUEG group, 33–67. Berlin: Language Science Press. DOI: 10.5281/zenodo.15775159.

Zeldes, Amir. 2018. *Multilayer corpus studies*. London: Routledge.

Zerbian, Sabine, Marlene Böttcher & Yulia Zuban. 2022. Prosody of contrastive adjectives in mono-and bilingual speakers of English and Russian: A corpus study. In *Proceedings 11th international conference on speech prosody. International speech communication association (ISCA) online archive*, 812–816. DOI: 10.21437/SpeechProsody.2022-165.

Zerbian, Sabine, Yulia Zuban & Martin Klotz. 2024. Intonational features of spontaneous narrations in monolingual and heritage Russian in the U.S.: An exploration of the RUEG corpus. *Languages* 9(1). DOI: 10.3390/languages9010002.

Zipser, Florian & Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on language resource and language technology standards, LREC 2010*. La Valette, Malta. https://hal.inria.fr/inria-00527799.