Corpus Annotation and Data Analysis (CAnDA)
Gauss Observatory, Göttingen
Liŋ'G VII, 19-30 September, 2022
Sum-School-22 (uni-goettingen.de)

Schedule for the CAnDA Workhop
Day One: 23 September (Friday)

Session 1 (chair: Lieke Hendriks)

10:30 – 11:00 **Vanessa Hagenschulte**
Emotionalization strategies in disaster reports: Challenges in the corpus compilation and analysis

11:00 – 11:15 Tea/Coffee Break

Session 2 (chair: Irene Pagliai)

11:15 – 11:45 **Gautam Ottur**
Arguments and events in verb series— insights across modalities

11:45 – 12:15 **Jian Ma**
Agent Prototypicality in Mandarin Chinese Passivisation: An Empirical Study

12:15 – 14:00 Lunch Break

Session 3 (chair: Marie Benzerrak)

14:00 – 14:30 **Prudence Pontbriand**
Object drop in Early Romance languages

14:30 – 15:00 **Albert González-Marín**
A Linguistic Corpus for Puerto Rican Spanish

15:00 – 17:00 Tea/Coffee + Poster Session

Posters

Abdelmagid Basyouny Sakr: On the nature of specialized collocations in Italian language of economics: a corpus-based study

Alexandra Chudar: Diminutiveness in Southern Hemisphere Englishes

Alina Sementsova: Prosodic features of impersonal sentences in Kazym Khanty: a corpus-based study

Anastasia-Milena Popovidou: Data collection methods for investigating language shift: the case of the Urum speaking community in Greece

Feras Saeed: Form-meaning mismatches in the noun phrase

Katharina Ronja Berking: The heroes of worlds past. Neomedieval narrative in the popular fantasy novel

Marie Benzerrak: Discovering an undescribed language: Corpus creation for Bokotá

Ming Liu: Orthographic competences of L1 Chinese students in L2 German: A learner corpus based multidimensional error analysis

Qiang Xia: A corpus-based study of turn-taking in online and face-to-face German conversation

17:00 – 18:00 **Invited Speaker: So Miyagawa** (chair: Elliott Lash)
Introduction to Cross-linguistic Syntactic Study through Universal Dependencies Treebanks and Syntax Parsing through BERT

Day Two: 24 September (Saturday)

Session 1 (chair: tba)

10:00 – 10:30 **Mirela Imamovic**
The statistical analysis of learners' L2 article choice

11:00 – 11:15 Tea/Coffee Break

Session 2 (chair: tba)

11:15 – 11:45 **Tamara Bassighini**
Word Formation in the Dialect of Tyrol: The Case of the Bavarian Verbal Prefix "der-"

11:45 – 12:15 **Romano Madaro**
Does the verb go "climbing"? Pinpointing OV/VO Alternation in the Northeastern Alps

12:15 – 14:00 Lunch Break

14:00 – 16:00 Tea/Coffee + Poster Session

Posters

Amereh Almossa: Discourse-pragmatic variation in Najdi Arabic
Andrea Mattichio: Low verbs, lower subjects: Word order phenomena in Old Italian
Lieke Hendriks: Finding discontinuous nominal constructions
Irene Pagliai: Preliminary steps in the creation of a cross-linguistic dataset for the study of ambiguity in idiomatic language
Luisa Gödeke: Viewpoint and perspective in non-fictional statements
Marie Christin Walch: Analysis and modelling of the denial of expectation contrast using the QUD framework (for now)
Wiebke Juliane Elter: Accommodation of Old Norse loan verbs in Middle English

Organized by:

Elliott Lash	elliottjamesfrick.lash@uni-goettingen.de
Irene Pagliai	irene.pagliai@uni-goettingen.de
Katja Friedewald	katja.friedewald@uni-goettingen.de
Lieke Hendriks	elisabeth.hendriks@uni-goettingen.de
Marco Coniglio	marco.coniglio@phil.uni-goettingen.de
Marie Benzerrak	marie.benzerrak@uni-goettingen.de

Email CAnDA ***ling.summerschool.2022@gmail.com***

The organizers thank Nicole Hockmann, Nermin Gürkan, Yana Strakatova, Matthias Kracht and Tanja Recke for their help with practical issues.

Abstracts

Invited Talks:

So Miyagawa:

Introduction to Cross-linguistic Syntactic Study through Universal Dependencies Treebanks and Syntax Parsing through BERT

This talk explores the possibility of cross-linguistic study using Universal Dependencies Treebanks. The Universal Dependencies (UD) is the uniform framework to annotate the syntactic information for as many languages as possible. So far, 228 languages and 130 treebanks are in the latest release on May 15, 2022. We can statically analyze the syntaxes of these languages and compare them through UD: e.g., word order, noun-adjective order, case alignment, the structure of relative clauses, etc. This talk will show the methods to do such analyses through UD. Furthermore, we train our automatic UD annotator using BERT, one of the deep-learning models, for any language.

Talks:

Vanessa Hagenschulte:

Emotionalization strategies in disaster reports: Challenges in the corpus compilation and analysis

Emotionalization is a mass media phenomenon that conflicts with the journalistic ideal of objectivity. But emotionalization does not necessarily occur through an explicit naming of emotions because it can be evoked by a diverse spectrum of linguistic triggers. While some studies from a modern synchronic perspective are already available, e-strategies from a diachronic perspective have not yet been systematically examined. This dissertation project is therefore dedicated to the development of linguistic emotionalization strategies in the British tabloid and quality press. Catastrophes such as reactor accidents are emotionally charged as such. This dissertation examines which emotionalization strategies are used in the relevant reporting on the events in Chernobyl (1986) and Fukushima (2011).

The analysis focuses on the following key questions: How do emotionalization strategies differ in tabloid and quality press? Which diachronic changes can be observed between 1986 and 2011? To answer these questions, it must be clarified which objectifiable methods can be used to capture the construct of emotionalization in terms of corpus linguistics, which also forms the main challenge of this dissertation. As a basis for the investigation, a corpus is compiled that includes approximately 400 articles from 1986 and 2011 from around nine newspapers. The methodological problem of this research lies in challenges in corpus compilation and techniques; this raises the question of transferring aspects of emotionalization into suitable investigation methods in the corpus. The corpus-based analysis is located in pragmatics and text linguistics. It combines quantitative and qualitative analysis approaches while using relevant software like AntConc.

The phenomenon of tabloidization suggests that there is an increase in emotionalising strategies not only in the tabloid press but also in quality newspapers. In addition, it should be checked whether target group-specific different emotions are described and evoked.

Gautam Ottur:

Arguments and events in verb series— insights across modalities

Serial verb constructions (SVCs) in sign languages are relatively understudied, and discussions of how event semantics and argument selection are entangled therein are crucially lacking. Although the literature identifies that co-eventive (tense-iconic) usages of verbs of transfer in series with lexical verbs often has argument-introducing behaviour (e.g. Aikhenvald 2006:25-26), the pervasiveness of this phenomenon remains unexplained. The present study contributes a formal comparison of these two dimensions in SVCs between a less-studied spoken language (Malayalam, Dravidian) and a sign

language (German Sign Language: DGS), which demonstrates that serializing behaviour remains relatively consistent across languages and modalities. I evaluate the behaviour of serial verbs of transfer in relation to cross-linguistic trends in event structure and referential uniqueness.

In both languages, usage of serial verbs of transfer exhibits an asymmetry dependent upon linearity, summarized in the table below.

Behaviour of 'give' and 'take' by linear order

Verb	V_1	V_2
'give'	*	Benefactive/recipient introduction.
'take'	Instrument introduction, event causation.	'take away,' 'gather'

These generalizations demonstrate the sensitivity of the verbs to iconic order within single-event predication; both 'give' and 'take' are preferred in argument-introducing contexts as V2 and V1 respectively. Give' is never co-eventive with the following verb due to its telic features, and 'take' only avoids triggering an event boundary as V2 by shedding its event-initiating interpretation (see Ramchand 2008 for further discussion). Modality appears to have virtually no impact on the order and function of verbs in SVCs. A preliminary analysis thus suggests that temporal iconicity conditions conventionalization of high-frequency verbs in SVCs, such as verbs of transfer or motion, and that single-event verb series are subject to the referential uniqueness constraint described by Bohnemeyer et al. (2007:519).

Jian Ma:
Agent Prototypicality in Mandarin Chinese Passivisation: An Empirical Study
Dowty's Agent Prototypicality shows a quantitative effect of agentive features: the prototype with the most agentive features is assumed to be the preferred candidate for linguistic constructions (Dowty 1991). By contrast, the effects of one specific role property are found in any amount of previous research and challenge Dowty's approach, e.g., [causation] in subject selection in English (Koenig & Davis 2001), [volition] in argument alignment in Experiencer-Object verbs (Verhoeven 2017) and [motion] in 3PL impersonal construction in Russian (Bunčić 2020). Such feature prioritization account can be explained by the role prominence feature approach (Himmelmann & Primus 2015).

My dissertation investigates these two approaches. An acceptability rating test (in progress) in Mandarin Chinese passivisation, i.e., bei-construction, is designed to figure out whether the Agent Prototypicality affects the passivisation. Six verb classes (five of them are adopted from a German test conducted by Kretzschmar et al. 2019) differing in the number of agentive features will be tested in active and passive voice in Mandarin Chinese. In this talk, I will present my experimental design (e.g., the identification of agentive features, a previous corpus test on the strong collocations of bei and the modification of Semantic Prosody in bei-construction) and, if available, also the experimental data.

Prudence Pontbriand:
Object drop in Early Romance languages
My research investigates null objects in the Early Romance languages. Null objects are attested in both Latin and some modern Romance languages, however they are differently licenced. In Latin, null objects were syntactically conditioned (i.e. they occurred in specific syntactic contexts, such as coordination, answers, non-finite clauses). In modern Romance languages (e.g. Brazilian Portuguese and Sursilvan), null objects are mostly semantically licensed (i.e. they are dependent on inherent properties of the object) (Pescarini, 2021: 101).

My research aims at answering the following questions: what are the conditions licencing null objects in Early Romance? Can we find any specific restrictions or triggers on null objects? Are these licencing conditions and restrictions the same across the early Romance languages?

I focus on two Early Romance languages: Old French and Old Italian (namely Old Tuscan). Studies of null objects have been carried out in both languages (see e.g. Luraghi, 1998; Egerland, 2003 for Old Italian and Arteaga, 1998; Donaldson, 2013; GGHF, 2020 for Old French).

A pilot corpus study yielded 77 examples of null objects in Old French and 59 examples of null objects in Old Italian. Overall, Old Italian and Old French have a similar profile of null objects. In both languages, null objects tend to be animate, often human, and specific. Similarly, both languages preferably omit 3rd person direct object.

However, referential null objects in Old Italian are a lot less frequent than in Old French, and occur in a more restricted context (namely only in coordination). Old French, on the other hand, shows instances of referential null objects in both coordination (within the same clause boundaries) and argument sharing contexts (i.e. in contexts in which the antecedent is in another clause).

Albert González-Marín:
A Linguistic Corpus for Puerto Rican Spanish

Spanish is one of the most commonly spoken languages worldwide, and yet little work has been put into studying its linguistic variety across Europe and Latin American until the past two decades. Even so, the work that has been done thus far has prioritized some varieties over other, due to more abundant resource materials, greater governmental support, etc.

One variety which has found itself neglected as compared to Iberian, Mexican, or even Cuban Spanish, for example, is Puerto Rican Spanish. This leaves researchers in an awkward position, as questions about the peculiarities of the variety spoken by the currently biggest figures in Spanish-speaking Pop Culture cannot be answered systematically.

As such, it would be of great utility to anyone looking to undergo such studies if an independent corpus specializing in Puerto Rican Spanish were available. As part of this project, we shall examine previous work into creating such resources, consider their failings, and propose new ways forward.

Hopefully, with such resources at hand, it would be easier to explain morpho-syntactic phenomena typical of Puerto Rican Spanish, which are often stereotyped as uneducated, mistaken, or otherwise incorrect by speakers of other varieties. Of course, an overview of such peculiarities is in order.

Mirela Imamovic:
The statistical analysis of learners' L2 article choice

Tamara Bassighini:
Word Formation in the Dialect of Tyrol: The Case of the Bavarian Verbal Prefix "der-"

Within the realm of German dialectology, but even more generally German linguistics, derivational morphology still constitutes a rather undiscovered field of research. With this in mind, my PhD project wants to shed light on a specific derivational morpheme prototypical of Bavarian varieties, which, due to its semantic and syntactic peculiarities, has been subject of discussion among German dialectologists since the second half of the last decade. Starting from the analysis of the derivational morpheme in the dialect of South Tyrol (Northern Italy), the intention of the project will be to investigate the influence of derivational morphology on the argument structure of the verb, considering possible effects on the verbal and inflectional phrasal domains (VP and IP respectively), as well as taking account of the grammatical categories of modality and aspect. The proposed talk will present the main points of the project and discuss some pertinent linguistic data gathered from historic and contemporary electronic corpora1.

The first part of the talk will introduce the inseparable verbal prefix “der-” (e.g. derfrieren, Germ. ‘erfrieren’ (“freeze to death”); deressen, Germ. ‘essen können’ (“manage to eat”)), explaining and exemplifying its most peculiar syntactic and semantic characteristics. The second part of the talk will then outline the aims of the project, as well as the theoretical and empirical tools intended to be used. In the third and central part of the presentation, approaches taken up to the present point of the ongoing research will be described and their results discussed. In particular, relevance will be given to data collected from two contemporary electronic corpora, the “DiDi Corpus of South Tyrolean Computer Mediated Communication” and the “Mobile Communication Database 2 (MoCoDa)”. Finally, the fourth and last part of the talk will be devoted to the discussion of the prospects and future steps of the project.

Romano Madaro:
Does the verb go "climbing"? Pinpointing OV/VO Alternation in the Northeastern Alps

In the field of linguistic variation, it has been already hypothesized (e.g. Gaeta, 2021; Gaeta and Seiler, 2021) that the alpine region should be considered as a linguistic union (= Sprachbund), given the patterns of convergence between languages/varieties belonging to different families (Romance/Germanic/Slavic), specifically in the domain of lexicon and morphology. In terms of syntactic features, recent studies (Rabanus and Tomaselli, 2017; Bidese and Tomaselli, 2021) have also provided further evidence in this sense: a fine-grained analysis of the (micro)parameters (in terms of Biberauer, 2010) correlated to the NSP-cluster between North-Italian dialects, Bavarian and Cimbrian/Mòcheno (two German-speaking islands in Trentino) displays different levels of variation in this specific area.

Applying the notion of “granularity” to the parameter [±OV], I intend to analyze the OV/VO alternation in subordinate sentences (e.g. dass-Sätze) by means of a comparison of the German varieties spoken in the Triveneto area. In addition to Cimbrian (now considered a VO-language: see Bidese et al., 2012; Poletto, Tomaselli, 2019), Mòcheno (OV/VO alternation due to pragmatic aspects: (Cognola, 2013), and Plodarisch (solid OV, but with initial switching signals: see Grewendorf et al., 2005.; Poletto, Tomaselli, 2019), I will also include some examples from a corpus of syntactic data of Timavese, whose structures have not been thoroughly investigated, to the best of my knowledge.

To do this, I will analyze the realization of V(P)R – a well attested phenomenon not only in these varieties, but also in West-Flemish and Züritüütsch – in the presence of the so-called "light" elements (e.g. object-pronouns, negation, Verbpartikeln). This will allow us to pinpoint the movement of the verb (=Vfin) towards the "higher" portion of the sentence, to trace a continuum of variation in these varieties, as well as to hypothesize a diagnostic in the passage from an Infl-final to an Infl-medial position, that is an initial switch between OV and VO-structure.

Posters:

Abdelmagid Basyouny Sakr:

On the nature of specialized collocations in Italian language of economics: a corpus-based study

Collocations are an important issue in modern linguistics. They exist in special languages as well as in general language. In my doctoral research I study the terminology of the language of economics used in Italian newspapers. Part of my work has to do with defining the status of the term in this domain on a normative level (both single or multiword terms). To this end we use the methodology of corpus linguistics and lexical semantics approach. Further, I deal with the topic of specialized collocations and their behavior based on data extracted the corpus I compiled. The Italian language of economics is characterized by containing a few technical terms and a high degree of collocability. Specialized collocations are defined as statistically significant co-occurrences of tokens or lexemes within the corpus. Our data shows that collocations come in various significant categories: from semantically non-idiomatic to full idioms. The data shows also that collocations could be a great source for extracting conceptual information related to: (1) identification of the conceptual metaphors of the domain, (2) concept identification, (3) concept systems, (4) particular meaning problems, and (5) terminological variation.

Alexandra Chudar:

Diminutiveness in Southern Hemisphere Englishes

Even though some scholars believe that the number of diminutives in English is quite low, several varieties are characterized as rather pro-diminutive. These are so-called Southern Hemisphere Englishes (Kachru 1992) formed at the end of the 18th century as a result of the second wave of British colonization. They include Australian, New Zealand, and South African varieties, with Australian English being most prone to diminutive formation (Simpson 2004, Kidd et al. 2011).

Corpus linguistics and sociolinguistics are both important and widely-developed areas of modern research. In my work, I use the combination of the two to discover the (ir)regularities of structural, semantic, and pragmatic variation of diminutives in the varieties of Southern Hemisphere Englishes. My study is based on lexicographic and corpus data, as well as employs some statistical methods to confirm the obtained results.

I focused on the variation of synthetic (doggie from dog), analytical (little dog), and inherent (whelp) diminutives in social context, where I discovered that variation is found in their structural and semantic characteristics, and in the diminutive "richness" of the varieties, while their functions are rather similar. The study has shown that of all the varieties under discussion, Australian and New Zealand Englishes share the majority of diminutives, while the South African English items are to some extent different, which is primarily attributed to the language contacts of the variety (influence of Afrikaans). In my report, I am going to provide more details on the cases of structural and semantic variation of diminutives in Australian English and the other varieties of the Southern Hemisphere, to focus on the reasons for this variation, as well as to dwell more on the usage of corpus data in the studies of lexical variation in English.

Alina Sementsova:

Prosodic features of impersonal sentences in Kazym Khanty: a corpus-based study

Almereh Almossa:

Discourse-pragmatic variation in Najdi Arabic

My research aims to investigate sociolinguistic variation and change in the formal encoding of the discourse-pragmatic features such as the epistemic expressions, general extenders, and intensifiers) in Najdi Arabic (NA). It seeks to explore the extent to which variation in the production of certain discourse-pragmatic variants correlates with functions and the speaker's age and gender. However, these features have not been studied in the context of spoken Arabic. In this study, the MA: 2ADRI: construction "I

DON'T KNOW" in Najdi Arabic (NA) will be examined in light of grammaticalisation, with an attempt to explore the functions of the phonetic variants in the interactional situation.

In order to achieve these goals, and due to the lack of available spoken corpus for NA, I built a corpus of spoken NA to be accessible for research in future1. The current dataset comprises about 18 hours of audio-recorded natural conversations with 60 native speakers of NA. Participants were equally stratified across age and gender. The recorded conversations were then segmented into turns and transcribed using ELAN software. To ensure consistency and to enhance the data for automatic annotation, the Orthography Convention for Dialectal Arabic CODA* is followed to transcribe all the conversations. CODA* was developed by Habash et al. (2018).

A total 700 tokens are examined, three different realisations are found: the full form *ma: 2adri*: [ma: ʔadri:], the semi-reduced form *ma: dri:* [ma: dri:] and the reduced form madri: [madri:]. The qualitative analysis demonstrates that MA: 2ADRI: construction performs multiple functions in the interpersonal and textual domains of discourse. The statistical analysis reveals a significant association between the three forms of the construction and the functions. While the full phonetic form is strongly connected with the literal meaning that expresses a literal meaning of lack of knowledge, the reduced forms are significantly associated with more discourse-pragmatic functions. Age appears to be a significant factor affecting the variation, whereas gender is not. Younger (aged between16 to 20) and adult speakers (aged between 30 to 40) are significantly more likely to use the reduced form, while the older speakers (aged between 55 to 70) are significantly more likely to use the full form. This can be interpreted as an indication of ongoing change led by the younger speakers towards greater use of the reduced form. Given the evidence of linguistic change in the reduced form including phonetic attrition, semantic bleaching and pragmatic strengthening, and its high frequency, the study suggests that the MA: 2ADRI: construction is undergoing grammaticalisation, with the *madri*: variant the most advanced form along the grammaticalisation cline (Traugott, 1989; Traugott & Trousdale, 2010).

Anastasia-Milena Popovidou:

Data collection methods for investigating language shift: the case of the Urum speaking community in Greece

The aim of the present study is to describe the relationship between linguistic and social factors in examining the language shift of Urum, an Anatolian Variety of Turkish spoken by Greeks from Georgia who currently reside in Greece. The ancestors of the Urum speakers came to Georgia from cities in Northeastern Anatolia (Kars, Erzurum) in the beginning of the 19th century and since then continued using the varieties of Anatolian Turkish in contact with the languages of the new environment, especially Russian, Georgian, and Armenian. When the Soviet Union collapsed in 1991, significant migrations of Greeks from the former Soviet territories took place, with the main destination being Greece. In this investigation, evidence will be presented to show how the significant social transformation of the Urum-speaking community after the migration to Greece (external factors) has led to changes in its ethnic self-perception, since the ethnolinguistic identity of Urum speakers is no longer used as an important distinguishing feature. In addition, this study will investigate different levels of individual's language competence, to understand additional factors that contribute to language shift. Broadly speaking, we will focus on Urum speakers who did not acquire a certain level of proficiency in the language, and they are unable to use it in socially significant ways, while, in our view, it will be a real challenge to pass that language on to children. Thus, the co-examination of the interaction between language-ethnic identity as well as of the pivotal role of other sociological factors that affect speaker's linguistic behavior and language competence as such, can shed light on the process of language shift and its structural consequences. Therefore, attention has been focused on certain processes of linguistic change that occur during the contact situations between Urum and Greek observed mainly in phonology, morphology, syntax, and lexicon and which may stand as indicators of language decay and language loss. In order to

collect the linguistic data, we use quantitative as well as qualitative research methods with different speakers of the community, from the fluent to the so-called semi-speakers so as to determine the impact of several external factors on speaker's behavior that lead to structural consequences and subsequently affect language maintenance itself.

Andrea Matticchio:
Low verbs, lower subjects: Word order phenomena in Old Italian
A vast body of research has proposed that Old Italian, along with other Old Romance varieties, displays a relaxed V2 grammar (Benincà, 2013; Poletto, 2014; Wolfe, 2019 a.o.). Main evidence in favour of this claim comes from the observation of instances of subject inversion that are no more possible in Modern Italian and the possibility to front an unspecified constituent. However, postulating a V2 grammar for Old Romance is not an uncontroversial assumption, and many scholars have tried to reject it, pointing out that the variety of word order options attested in the Old Romance left periphery is not compatible with a V2 constraint, and that alternatives are possible (Kaiser, 2002; Martins, 2019; Mensching, 2012; Rinke and Elsig, 2010 a.o.). The problem is particularly salient for Old Italian, where V1 and V3+ orders are robustly attested and do not show any obvious signs of grammatical constraints. On the other hand, some studies have suggested that different word order configurations in Old Romance are associated with different information structures (Ciconte, 2018; Rinke & Meisel, 2009), independently of a Germanic-like V2 constraint.

In this work, I will claim that assuming that word order in Old Italian varieties is conditioned by a V2 grammar is descriptively inadequate, and may lead to undesirable conclusions about the nature of V2 languages in general. Instead, I will discuss whether information structure and discourse may be more useful notions to predict the frequency of certain word order patterns. With a collection of data from Old Italian and an annotation for some information-structural parameters, I will look for correlations between the position of the subject and its interpretation, testing in particular whether one specific post-verbal position is preferred when the subject escapes a focal interpretation.

Lieke Hendriks:
Finding discontinuous nominal constructions
Discontinuous nominals, or split Noun Phrases (NP), are constructions in which one constituent has presumably been split into two parts, e.g. a topic and a remnant. The examples in (1) and (2) illustrate this construction. In sentence (1), the presumed underlying construction "drie boeken" has been split up into topic "boeken" and remnant "drie", and in (2) "geen koeien" has been split up into topic "koeien" and remnant "geen".

(1) Boeken heb ik drie.
Books have I three
"As for books, I have three."
(2) Koeien heb ik geen.
Cows have I no
"As for cows, I have none."

The search for split NP constructions as in (1) and (2) in corpora may not be an impossible task, but it is certainly not an easy one. One issue concerns the compilation of a corpus that one needs to factor in. The corpus research in my project so far has looked at split NPs in CGN ("Corpus Gesproken Nederlands" Corpus Spoken Dutch) and SoNaR (Dutch reference corpus) using two online platform tools: GrETEL and OpenSoNaR. These two online search engines use different mechanisms to search through the data in CGN and SoNaR. The first one, GrETEL, parses the data into tree structures and lets you search through the data with example-based queries, whereas the second one, OpenSoNaR, tags the data and lets you search the data using Corpus Query Language (CQL).

Another issue concerns the setup of a query. So far, two approaches have been used within this project to look for split NPs as in (1) and (2). The first approach concerns looking for a construction that concerns both split parts: [Topic […] Remnant]. The second approach zooms in on the remnant and tries to look for stranded determiners, adjectives or quantifiers: [Remnant […]]. Depending on the compilation of the corpus, either approach could prove useful.

Feras Saeed:
Form-meaning mismatches in the noun phrase

My research focuses on nominal mismatches in multiple nominal structures that lack a 1:1 transparent mapping between form and meaning. In particular, I look at certain nominal structures where inflection on the noun and its modifiers does not seem to contribute to the meaning of the noun phrase. To this end, I investigate several instances of nominal mismatches in number, gender, definiteness and case between the noun and two nominal modifiers: adjectives and numerals.

In adjective-noun structures, I look at why postnominal adjectives display mismatches in number and gender in the presence of certain classes of nouns, and why they fail to inflect for the these features altogether when they appear in a prenominal position. I also look at why certain features, e.g. case and definiteness, seem to escape certain restrictions on nominal inflection in these structures.

In numeral-noun structures, I look at how certain numerals can carry a gender marker different from the one on the noun, and why nouns fail to inflect for plural number in the presence of certain numerals. In addition, I look at how/why most of these mismatches are not triggered in these structures when the numeral is postnominal.

Irene Pagliai:
Preliminary steps in the creation of a cross-linguistic dataset for the study of ambiguity in idiomatic language

Idioms are multiword expressions (MWEs) whose overall meaning is not derivable from the sum of the literal meanings of the internal constituents (e.g. *kick the bucket* means "to die"; v. Sailer, 2020, 2021). Idioms are therefore naturally characterized by a mismatch between form and meaning. However, in some contexts the compositional meaning of an idiom may coexist with the figurative one, thus creating ambiguity (Wagner, 2020). See the examples below:

1. But, if I do end up on some opiate,! *the bucket has been kicked* and leaking significantly

   [pilotsofamerica.com]

2. [...] everyone after Katrina was opening their home to all sorts of destitute animals – cats and dogs – *it was raining them* for a while.

   [enTenTen15]

Idioms form a very heterogeneous group (Bizzoni et al. 2018), and are distinguished from each other on the basis of some dimensions of variation. It is common claim that only highly decomposable, transparent and literally plausible idioms can occur in ambiguous contexts (Vulchanova, Vulchanov, 2018; Haagsma, 2020; Wagner, 2020). However, *kick the bucket* (ex. 1) is the prototype of the non-decomposable and non-transparent idiom (Nunberg et al. 1994), while *it's raining cats and dogs* (ex. 2) is not literally plausible. This raises the question of the relationship between idiom-internal features and ambiguous uses in context: how much do idiom-specific features affect the production and understanding of

ambiguous occurrences? Is it possible to identify categories of idiomatic ambiguity connected to the idiom internal features?
To answer these questions, the first part of my PhD project consists of creating a new cross-linguistic (English and Italian) lexicon of idioms accompanied by ratings regarding decomposability, transparency and literal plausibility assigned by native speakers. In doing so, the project partially fits a well-established research strand in psycholinguistics (see among others Titone, Connine, 1994; Tabossi et al. 2011; Bulkes, Tanner, 2017; Hubers et al. 2019); motivations, challenges and new insights encountered so far will be illustrated.

Katharina Ronja Berking:
The heroes of worlds past. Neomedieval narrative in the popular fantasy novel.
My dissertation deals with questions of the popularity of the Middle Ages in popular fantasy literature. Discursively, it will be explored which aesthetic and narrative features popular fantasy literature exhibits that can be summarized into a neomedieval grammar. This will be investigated using the example of heroic characters, which - according to the thesis - follow a medieval staging in the corpus. In order to be able to do justice to the extensive amount of text in the corpus of popular fantasy literature (including, for example, The Lord of the Rings by J.R.R. Tolkien, The Witcher by Andrzej Sapkowksi, and A Song of Ice and Fire by George R. R. Martin, to name only the very big ones), corpus linguistic methods will be used to analyze the amount of data. Corpus linguistic methods are particularly suitable for my research because they bring unconscious structures to light in the sense of a discursive approach. These allow statements to be made about the discursive knowledge anchored in contemporary literature, a knowledge that superficially stages knowledge about the Middle Ages - so the thesis goes - updates it, and at a deeper level negotiates modern discourses within it. This neomedieval knowledge, in turn, allows for statements about the popular staging of the Middle Ages in contemporary literature as part of our popular culture. The last months were mainly used to compile a corpus of texts from popular fantasy literature. The aim was to verify popularity via quantitative methods and to distinguish fantasy literature that negotiates medieval discourses from other fantasy literature. Often the term high fantasy is used in this context, but rather the term epic fantasy has proven to be viable. At this stage, I am evaluating which methods seem appropriate for analyzing the corpus of fantasy literature to my research questions.

Luisa Gödeke:
Viewpoint and perspective in non-fictional statements
Fictional utterances are not assertions. Instead, they are assumed to be invitations to imagine things, scenes and events (Currie 1990). Nevertheless, a fictional work might also consist of statements that do not primarily serve to build up the fictional (i.e. imaginative) world (Searle 1975; Konrad 2014). Instead, so called non-fictional statements can be considered to be true in the real world: Either because they provide background information on real world events and persons or because they present states of universal validity. One of the best-known examples of this sort of utterance is the first sentence of Lev N. Tolstoij's Anna Karenina:

(1) Every happy family is alike, every unhappy family is unhappy in its own way.

We annotate non-fictional statements based on narratological guidelines in German fictional texts from 1600–1950. Currently, we have 19 texts partly annotated with 107.700 tokens in sum and 15.800 tokens of non-fictional readings (14% frequency). The findings are syntactically and semantically heterogeneous. Yet, assuming nonfictional statements to be a mode of discourse, I am particularly interested in identifying similarities of view point expressions, including deictics and tense and also presumably 'neutral' expressions such as negation and determiners (Dancygier 2012b). In my presentation, I will present a first analysis of non-fictional readings as driven by viewpoint shifts and

perspective markers. Thereby I argue that point of view in narrative impacts even the fictionality dimension.

Marie Benzerrak:
Discovering an undescribed language: Corpus creation for Bokotá

My thesis project aims to describe the Bokota language, a language of the Chibchan family spoken in Panama, and to analyse in detail the consequences of information structure on word order.

Bokota, as well as the other chibchan languages, is an SOV language (see Chamoreau 2017, 2018, 2019 for Pesh, Verhoeven 2012 for Cabecar, Herrera 2017 for Malecu, Bajorat 2014 for Ika). Having very little morphology, Bokota allows for variation in word order and we can observe SVO and OSV constructions. These modifications can be triggered by information structure, namely the marking of topics or focuses. (1a) shows the unmarked word order. (1b) explicits a Ofocus construction with an OSV order. If this order is accepted in other languages of the family, different structures are generally observed. In Pesh, for example, the OSV order is more often used in OSfocV constructions (Chamoreau 2017, 2018, 2019). Thus, within the same family, two different pragmatic strategies can give the same surface order. The analysis of less described languages may in the future open up new perspectives for analysis. The SVO order is possible, as in (1c), but it occurs rarely and under restricted contexts (Benzerrak 2018). (1d) has not been attested in Bokota yet but the OVS order is allowed in most chibchan languages. For example, Cabécar, a language from the same branch as Bokota (Isthmian branch) respect an Absolutive-V constraint that allows only two word orders regarding the transitive verb and its arguments: SOV and OVS, the object being restricted to the direct left position of the verb (Quesada 2007).

(1) Bokota, Benzerrak (2022, pers. data)

a. Koi kle bligda gut-e d. ? bligda gute koi kle
   chicken PRS food eat-PRS
   “The chicken is eating the food.”
b. bligda koi kle gut-e e. *gute koi kle bligda
   food chicken PRS eat-PRS
   THE FOOD, the chicken is eating (it).”
c. koi kle gut-e bligda f. *gute bligda koi kle
   chicken PRS eat-PRS food
   “The chicken is eating the food.”

The facts in (1) are established with a questionnaire that have been conducted with native speakers of Bokota. The questionnaire was analysing the factors FOCUS POSITION (Sfoc and Ofoc) and WORD ORDER (SOV, OSV, SVO and OVS) and consisted of a first question and answer task, and a second acceptability judgement task. The challenge is to compare the intuitions from these experimental data with facts from spontaneous speech production.

In my presentation, I will present the frequencies of the configurations in (1) in a text sample of five narrative texts, collected in the field with six native speakers of Bokota. It is a sample of 28 texts (one speaker recorded only three narratives) and one hour of recording. The themes of the texts are previously established and are the following: cooking recipe, history of ancestors, path from one place to another, construction of the traditional house, and comparison between Bokota and Bugle (a neighbouring community). They are reproductible in all chibchan languages and can be added in the Pesh corpus (Chamoreau, available on ELAR Archives), the Cabécar corpus (Verhoeven, available at the Language Archive), and the Maleku corpus (Herrera, available on ELAR Archives).

The data is analysed in Elan and Flex. The first software allows us to annotate any layers, in particular an 'information structure' layer. The second software is useful to automate the morphological transcription of the narratives.

From this sample the main questions are:

- Are the acceptability contrasts in (1) replicated in corpus data?
- Does the contextual information of corpora (opposed to experimental data) contributes to our understanding of the contextual restrictions that determine the possibility to use (or not) a certain word order?

Marie Christine Walch:
Analysis and modelling of the denial of expectation contrast using the QUD framework (for now)
Various contrast markers such as the German discourse connectives 'aber' (but) or 'obwohl' (although) connect two propositions that contradict each other. In particular, they can trigger a so-called denial of expectation, the strongest type of contrast according to Hobbs (1985). This occurs when the expected value of the background assumption, which is based on domain knowledge, personalexperience or social behavioural norms, is not adhered to or even rejected. In the QUD Gen project, a tool was used to extract and annotate sentences with different contrast markers from a corpus of vehicle data using the QUD framework, such as: "During the brisk but not heated drive through the Provençal Alps, the on-board computer then showed 17.7 litres after all." In this example, the expectation is not met that the vehicle's low consumption will be maintained even during correspondingly demanding driving. An already developed module predicts contrasts based on discourse markers in the text (Langner 2020). An additional module is in the planning stage that will autonomously generate sentences with contrasts based on the captured vehicle data from the underlying corpus.

My poster presentation will focus on defining the specific requirements for the planned module and also its limitations, which have already resulted from the manual annotation of the contrast sentences. The biggest challenge seems to be the generation and estimation of expected values. Some contrasting statements cannot be traced back to numerical data, or only very vaguely. They are often based on abstract or subjective adverbials and adjectives such as 'feinfühlig' (sensitive), which can only be converted into an expected value with great effort, if at all. In the previous manual annotation of the present corpus, it also became apparent that expectation values rather corresponded to expectation intervals that do not contain the contrastive value. These expectation intervals can only partly be derived as a function of existing technical data of the vehicle itself. In some cases, this value is derived in a generalised way from the vehicle class (such as small car or SUV) of the respective vehicle. Finally, the functionality of the module needs to be evaluated in an acceptance study. For this purpose, test subjects are presented with generated sentences whose naturalness is to be rated on a 5-point Likert scale.

Ming Liu:
Orthographic competences of L1 Chinese students in L2 German: A learner corpus based multidimensional error analysis
Although learners with L1 Chinese represent one of the largest learner groups of German as a foreign language, there are currently no freely available large-scale Chinese learner German corpora available, either in German-speaking countries or in China. In the context of my doctoral thesis" German orthography acquisition of Chinese learners - a corpus-based study on orthographic errors", the learner corpus DeChiLKo will be created. In the context of the CAnDA summer school, I would like to give a poster presentation about the structure, the current state of development of the corpus, and the first results of the data analysis.

The learner corpus is composed of two sub-corpora - DeChiLKo examination corpus and the DeChiLKo acquisition corpus. The examination corpus contains 195 dictations under the PGG examination condition in the "dictation" part (Examination Center 2013), which are written by 195 Chinese German students from 20 different Chinese universities in 2017 and 2019. The acquisition

corpus contains dictation exercises in class and exams from 5 Chinese German students in their first three semesters. In addition to the learner texts, the dictation corpus also contains the transcription of the dictation audio file. Metadata and speaker metadata are collected for all texts and stored according to the standard proposed by Granger/Paquot (2017).

Analogous to the other written learner corpora, all learner texts in DeChiLKo are machine-tagged for word types, lemmas, and sentence spans after tokenization by calling the TreeTagger internally in EXMARaLDA (Dulko) (Nolda 2019). As in the Falko corpus and Kobalt-DaF corpus (Reznicek et al. 2012, Hirschmann/Nolda 2019), the original text of dictation is manually inserted as a target hypothesis (Lüdeling/Hirschmann 2015), which is likewise also automatically annotated with part of speech, lemmas, and sentence spans. Differences between tokens of the target hypothesis and the learner text are automatically detected and marked with tags such as INS, CHA, SPLIT, etc.

The special feature and at the same time the focus of DeChiLKo is that the orthographic deviations are annotated at each position concerning different orthographic dimensions in the German language in the respective error levels. The tokens are segmented by graphemes, syllable structures, and morphemic structures and marked by using the predefined error tagset. The whole annotation process is done by internal retrieval of transformation scenarios in EXMARaLDA (Dulko), but most of the scenarios have been adapted to the research needs to facilitate the annotation process.

DeChiLKo is currently in the annotation phase, so far 200 learner texts with 9,272 tokens have been fully annotated. After the conversion of the EXMARaLDA files in makedulko (Nolda 2021), the entire DeChiLKo corpus will be available for search and analysis in the form of an ANNIS corpus.

Qiang Xia:
A corpus-based study of turn-taking in online and face-to-face German conversation
Short latency in video-call can affect the smooth turn-taking in conversation. This study investigated turn transition behaviors in dyadic conversation over Zoom and in a face-to-face situation from a temporal aspect, using dialogues collected in the Berlin Dialogue Corpus (Belz et al., 2021), where 20 native German-speaking participants were asked to finish a spot-the-difference task in dyad within 10 minutes, respectively two tasks in the two conversation conditions. An annotating scheme regarding turn-taking phenomena (e.g. speech turns, backchannels, gaps, overlaps, see Sacks et al., 1974; Heldner and Edlund, 2010) was developed for the purpose of the study. Methodological issues regarding the annotation and statistical treatment of the duration are discussed. Inter-speaker gaps averaged 395 ms in co-present conversations, but gaps for the same dyads over Zoom averaged 540 ms, leading to an increase in turn overlaps and interruptions in video-based dialogues. Compared with the co-present situation, online conversations have a higher speaker change frequency and receive far more back-channeling responses, suggesting speakers use a different turn-taking pattern in video-based conversation.

Wiebke Julianne Elter:
Accommodation of Old Norse loan verbs in Middle English
The basic vocabulary of English (Durkin 2014; Grant 2009) is rife with loans, like Old Norse (ON) give. The number and nature of borrowings resulting from a contact situation depend on the contact's intensity (Thomason & Kaufman 1988), but also on the morphological complexity of the borrowable categories (Matras 2009: 175f.). From this follows that linguistic closeness of the languages in contact could favour borrowing of complexer categories (Winford 2003: 51ff.; cf. Johanson 2002). When entering a language, loan words are integrated grammatically into the recipient language (Muysken 2000). For verbs, the structural implications of loan integration are often focused on less in models of borrowability and loan integration (cf. Matras 2007; Thomason & Kaufman 1988) or are operationalized as a constraint on borrowing (cf. Winford 2003). Wohlgemuth (2009) has found for those formal aspects that direct insertion is the most frequent accommodation strategy cross-linguistically. Hence, inflection should not

constrain loan verb integration. However, recent research shows that, even under direct insertion, loan verbs are subject to constraints and enter some usage categories more readily than others (De Smet 2014; Shaw & De Smet 2022). Considering the varying nature of linguistic contacts this work investigates whether these 'accommodation biases' hold in the typologically and lexically close contact between ON and English. A corpus study on ON loan verbs in Middle English (ME) compares their overall usage and the course of their structural integration to a set of native English verbs. Data are extracted from the Penn-Helsinki Parsed Corpus of Middle English (Kroch & Taylor 2000). Quantitative analyses gage the impact of etymology and finiteness on the integration of ON loan verbs in ME. If the analysis shows that accommodation biases are not significant for these verbs, one may conclude that biases are less prominent in contact between typologically closer languages.