Workshop: Second Austrian Meeting on Digital Linguistics / Zweites Österreichisches Treffen zur Digitalen Linguistik

Tanja Wissik (Österreichische Akademie der Wissenschaften), Andreas Baumann (Universität Wien), Julia Neidhardt (TU Wien), Claudia Posch, Gerhard Rampl (Universität Innsbruck)

Digitale Linguistik ist ein wachsendes interdisziplinäres Feld an der Schnittstelle von Linguistik, Informationstechnologie und Sozialwissenschaften, was sich auch in neuen Projekten, neuen Publikationsreihen und Lehrveranstaltungen zeigt. Ein zentraler Punkt der digitalen Linguistik sind die Sprachdaten, d.h. digitale Artefakte, die die menschliche Sprache als Ausdrucksform verwenden. Die Bandbreite dieser Sprachdaten reicht von Social Media Inhalten, über Parlamentsprotokolle und Zeitungen bis hin zu mittelalterlichen Manuskripten. Solche Daten werden unter anderem verarbeitet, annotiert, analysiert, kuratiert, geteilt, archiviert und wiederverwendet. Daher erstreckt sich das Themenspektrum von der Erstellung von z.B. digitalen Sprachressourcen (Corpora, Wörterbücher etc.) und deren Analyse (z.B. automatisierte Erkennung semantischen Wandels, Sentiment- und Emotionsanalyse etc.) über die Verwendung von Standards und Forschungsinfrastrukturen bis hin zu Methoden der Langzeitarchivierung oder Nachnutzung dieser Sprachdaten.

Die Bandbreite der Forschungsaktivitäten in diesem Bereich in Österreich wurde bereits beim (ersten) Österreichischen Treffen zur Digitalen Linguistik: Rezente Entwicklungen in Österreich bzw. bei den bisherigen Österreichischen Treffen zu Sentimentinferenz (ÖTSI 2021, 2023) sichtbar, bei denen insgesamt 37 Forschende aus unterschiedlichen österreichischen und internationalen Forschungseinrichtungen ihre Projekte präsentiert haben.

Der diesjährige Workshop "**Zweites Österreichisches Treffen zur Digitalen Linguistik"** versteht sich als Fortsetzung dieser Workshops und hat auch wieder zum Ziel die rezenten Entwicklungen in der Forschungslandschaft in Österreich aufzugreifen, und unterschiedliche Projekte aus der österreichischen Forschungslandschaft, welche mit oder an Methoden der digitalen Linguistik arbeiten, und die darin involvierten Wissenschaftler*innen zu vernetzen. Es sollen dadurch (1) methodologische Erkenntnisse ausgetauscht werden und (2) Synergien durch das gegenseitige Verfügbarmachen von digitalen Sprachressourcen entstehen, auch im Rahmen der Forschungsinfrastruktur CLARIAH-AT.

New ways to identify phraseological patterns: between statistics and AI

Dominique Longrée (Université de Liège) Laurent Vanni (CNRS – Université Côte d'Azur)

In this paper, we will first briefly review the different types of patterns considered as "phraseological" by the linguists. We will then precise some difficulties their automatic detection meet, and we will evaluate some pattern detection techniques (combining data mining and statistics), in order to assess their performance, advantages and disadvantages. We will finally explore to which extend the use of HyperDeep, an AI tool, may prove useful, or even indispensable, in this field. The automatic detection of phraseological patterns meets various difficulties when the notion is extended to non-totally fixed patterns (Longrée & Mellet, 2013): unlike repeated segments (Salem, 1986), verbal tense sequences (Longree & Luong, 2003) or syntactic motifs (Mellet & Longrée, 2009), these non-totally fixed patterns (Fillmoreet alii 1988 ; Sinclair, 1991: 111-112 ; Gledhill & Frath 2007) consist in multidimensional phraseological patterns made up of items of several natures (forms, lemmas, grammatical categories, etc.) and allowing various variations : lexical or morphological variations, permutations, presence or absence of elements, spaces between pattern elements. Each of these variations is a new challenge for automated detection based on a sequential search. A tricky parameter is the frequency of the patterns : repetitions are necessary to insure the memorization of the patterns (Lavigne et alii 2016; Longrée, Mellet & Lavigne, 2019), but a high frequency of a pattern does not always mean that the pattern is phraseological: syntactic patterns are highly frequent, but cannot of course be considered as "phraseological unit"; some of the repeated patterns are not fixed as real "phraseological units", but function as "motifs" with a textual function, while others function both as phraseological units and "textual motifs" (Longrée & Mellet 2018). Nevertheless, several tools have been proposed since the beginning of the century in order to detect and distinguish phraseological units and "textual motifs" (see Ganascia

2001). We will test some of these tools: SDMC (Sequential Data Mining under Constraints; https://sdmc.greyc.fr/index.php; Cellier *et alii* 2012; Quiniou et alii 2012), Le lexicoscope (Kraif 2016; 2019) and Hyperbase (https://hyperbase.unice.fr/ (Vanni 2024). In order to assess the advantages and disadvantages of each of these methods, we will apply them to a corpus of Latin texts processed using LASLA methods. We will finally compare the "linguistic patterns" we extracted with those the new software HyperDeep (based on CNN and Transformers; Vanni *et alii* 2018, 2023; 2024) identifies in the same corpus and shows the added-value of this method.

Claudia Mattes (Universität Wien)

Das gehören-Passiv. Korpuslinguistische Untersuchung einer analytischen Konstruktion. Erfahrungen und Herausforderungen.

Die Grammatikalisierung des gehören-Passivs habe ich bereits in meiner Masterarbeit erforscht (siehe Mattes 2024), ein facettenreiches Thema, wofür das Austrian Media Corpus (amc) als Datengrundlage verwendet wurde. Die Erfahrungen daraus sollen nun in eine Dissertation einfließen, deren umfangreicherer Rahmen für einen detailliertere ausgeweitete Untersuchung auf unterschiedlichen Ebenen genutzt werden soll. Im Forum des Workshops würde ich die bisherigen Ergebnisse und die weitere Vorgehensweise vorstellen sowie die Gelegenheit wahrnehmen, um über das konkrete Phänomen hinausgehende Aspekte der Untersuchung und Anliegen digitaler Linguistik zu diskutieren. In der Forschungsliteratur wurden zur Erforschung des gehören-Passivs entweder einzelne Beispiele angeführt (siehe Szatmári 2002) oder bestimmte Kombinationen mit CQL abgefragt (siehe Stathi 2010; Lasch 2016), wobei schwierigere Fälle – wenn – nur am Rande beleuchtet wurden. Um die Konstruktion möglichst umfassend fassen zu können, habe ich einen Zugang gewählt, der eine NLP-Pipeline miteinschließt. Die hierbei auftretenden Herausforderungen, bedingt durch die Nutzungsbedingungen des amc oder die Annotation durch unterschiedliche Parser, wurden mithilfe verschiedener Mittel bewältigt: Unter anderem wurde zusätzlich zum quantitativen Korpus von ca. 42.000 Belegen ein qualitatives Sample manuell validiert und unter gewissen Gesichtspunkten annotiert, um die automatisiert extrahierten Hits zu kontextualisieren. Ein Anliegen ist unter anderem, die Erkenntnisse aus der bisherigen Untersuchung und die erweitert annotierten Daten für die weiteren Analysen zu nutzen; eventuell in Form eines eigenen Parsers oder einer anderweitigen Form der Klassifikation. Die Konstruktion gehören + Perfektpartizip soll in weiteren Kontexten untersucht werden, denn die Ergebnisse legen nahe, dass es sich um ein tendenziell mündliches, eventuell aus dem standardferneren Gebrauch stammendes Phänomen handelt (vgl. Mattes 2024: 116), entgegen der Schrift- und Standardsprachlichkeit, die grundlegend das amc prägt. Zudem ist nach wie vor die Entstehung des gehören-Passivs nicht eindeutig geklärt: Eine diachrone Perspektive sowie der Abgleich mit anderen gehören -Konstruktionen wie Reflexiv (Szatmári 2002) und Qualitativ (Lasch 2018) sollen hierfür fokussiert werden. Mehrere Korpora – schriftliche und auditive Daten, auf der vertikalen Achse zwischen Dialekt und Standard, rezente und historische Quellen– bedeuten unterschiedliche Herausforderungen, auch mit den Mengen der Daten, die extrahiert werden dürfen und können, um die Filterung mithilfe NLP zu bewerkstelligen und ausreichend Belege zu sammeln. Eine Auswahl soll dahingehend mitsamt Problemlösungsstrategien zusammengestellt werden. Der Wunsch besteht, auf Basis der bisherigen Erkenntnisse einen Workflow zu entwickeln, der für die Untersuchung dieser und ähnlicher Passivgrammatikalisierungen funktioniert, aber auch für andere Konstruktionen, die, eben aufgrund ihrer Form– bspw. Diskonnektivität – oder weil sie außerhalb der Standardsprache bzw. Norm liegen, bislang weniger Aufmerksamkeit bekommen haben.

Markus Pluschkovits (Österreichische Akademie der Wissenschaften, Universität Wien)

Progressive Aspektualität des Deutschen quantitativ untersuchen: Formen, Funktionen und Variation

In jeder natürlichen Sprache kann die Vorstellung kommuniziert werden, dass ein Vorgang sich dynamisch entfaltet. Diese Vorstellung, die ihren Ursprung in der Ereigniskonzeption von Sprecher:innen hat und zu der Domäne der Semantik gehört, bezeichnet man als progressive Aspektualität (siehe Mair 2012: 803). Besitzt eine natürliche Sprache eine formalisierte, grammatische Kategorie, die eine solche

Aspektualität markiert (beispielsweise durch ein *-ing* Suffix am Verb, wie im Englischen), so spricht man vom Progressivaspekt.
Im gegenwärtigen Deutsch gibt es keine vollständig grammatikalisierte Kategorie, um sich dynamisch entfaltende Vorgänge auszudrücken und dementsprechend auch keine grammatische Kategorie des Aspekt (siehe Zifonun / Hoffmann / Strecker 2011: 1861). Dies ändert jedoch nichts daran, dass progressive Aspektualität von Sprecher:innen ausgedrückt wird – in Ermangelung einer grammatischen Kategorie dafür teilweise mit sehr heterogenen Strategien. Die Untersuchung dieser Strategien mit digitalen, quantitativen Methoden und in einem konstruktionsgrammatischen Framework bildet den Rahmen meines Dissertationsprojekts.
Als Exemplar dieser diversen Strategien (die teilweise besser erforschte Konstruktionen wie den *am*-Progressiv umfassen, aber auch marginalere Konstruktionen wie periphrastische *tun*-Konstruktionen) soll für den Beitrag die pseudokoordinierten Positionierungsverbkonstruktion dienen – also *sitzen, stehen* oder *legen* mit einem Vollverb (vgl. Proske 2023), wie beispielsweise *du sitzt da und siehst dir beim Verfaulen zu* (Welt 2013). Der Status dieser Konstruktion als Ausdruck progressiver Aspektualität ist dabei umstritten, Ebert (2000: 607) beispielsweise behauptet, dass diese Konstruktion in einer progressiven Aspektualität im Deutschen nicht existiert, gegenteilige Befunde liefert Proske (2023).
Fokus des Beitrags soll dabei der Workflow zur Entschlüsselung dieser Konstruktion mit quantitativen, digitalen Methoden sein, als Stellvertreterin für alle im Rahmen des Dissertationsprojekts erforschten Konstruktionen. Dabei soll gezeigt werden, wie ein Korpus eines vergleichsweise seltenen syntaktischen Phänomens aus der diversen Korpuslandschaft für die deutsche Sprache aufgebaut werden kann. Besonderes Augenmerk soll dabei auf Repräsentativität der verschiedenen Medien (mündlich und schriftlich) und Register gelegt werden. Weiters sollen verschiedene Möglichkeiten der quantitativen, datengeleiteten Analyse anhand des Korpus aufgezeigt werden, wie beispielsweise die Sentimentanalyse als Möglichkeit, etwaige subjektive Funktionen der Konstruktion (im Sinne Traugotts 1989) aufzuzeigen (siehe auch Proske 2023) oder
eine statistische Analyse der regionalen und medialen Verbreitung der Konstruktion für eine Verortung am vertikalen und horizontalen Varietätenspektrum (d.h., wird diese Konstruktion eher schriftsprachlich oder gesprochensprachlich assoziiert, und beschränkt sie sich auf spezifische Dialekträume des Deutschen). Erste Ergebnisse zu diesen Analysen sollen den praktischen Teil des Beitrags abrunden.

Nikola Dobric, Ulrike Krieg-Holz, Luca Melchior (Universität Klagenfurt)

The Corpus of Austrian German – Construction of a 'national' language repository for research and socio-cultural purposes

Contemporary research into language and the society that uses and had used it is based on the electronically available language material found in the form of language corpora. Standing out among the many types of corpora available today are those known as, for a lack of a better term, 'national'. What differentiates the so-called 'national' corpora is the purpose for which they are constructed and the sources that are used to build them. The purpose is one of an all-encompassing repository of one language considered as a sociolinguistic whole. They are not only intended for a variety of types of linguistic research (from lexicography to language policy) but also for gaining socio- cultural insights into the society represented by such a general monitoring (i.e., all- encompassing and constantly updated) corpus. In order to serve such a demanding purpose, the sources for the construction of a 'national' corpus imply as high level of representativeness of one language as practically possible. Representativeness here implies a wide range of varieties and sources, as well as both a synchronic and diachronic focus. To this end, 'national' corpora are usually positioned as important strategic projects, designed to represent ongoing archives of one language in all of its domains and uses. Our project proposes the creation of a 'national' corpus of this kind for Austrian German. We outline the need for it, the potential structure, the resources necessary to initiate its compilation and continuation, the existing corpora of Austrian German which could be pooled under the aegis of it, and a tentative road map for creating it.

Pia Lehecka (University of Edinburgh)

Exploring a Historical Phonological Corpus: The case of long back vowels in Older Scots

Over the past three decades, corpora have been increasingly compiled to quantitatively study the diachronic development of languages. However, the vast majority of corpora parse historical texts on a morphological, syntactic, semantic or lexical level, while phonological historical corpora are notably absent (Molineaux et al. 2023). As a historical phonological corpus, the *From Inglis to Scots* (FITS) corpus combines the advantages of quantitative corpus linguistics with time-tested methods of historical phonology and allows a phonological investigation of historical sound changes.

This paper demonstrates the applications of FITS — a grapho-phonologically parsed corpus of the earliest attested stages of Scots —by tracing the spellings of Older Scots long back vowels /oː/, /ɔː/, and /uː/ between 1380 and 1500, and reconstructing the sound changes that underly them. Studying these long back vowels can provide novel insights into the broader category of changes known as 'The Great Vowel Shift', both in Scots and in its sister language, English. Lack of raising in the Scots back vowels has, indeed, been given as evidence for the GVS being a 'push chain' (Luick 1921, Lass 2000) in both languages. By comparing the sound systems of the main source language of Older Scots (1380-1700), Old English (500-1100), and its descendant, Modern Scots (from 1700), FITS allows us a first, broadly quantitative approach to these sound changes, and many others.

Corpus design

To create FITS, c1,100 root morphemes were phonologically reconstructed based on the *littera* theory, which argues that spelling variants and outliers found in a pre-standard language can indicate sound changes. Each phonological reconstruction triangulates between spelling, sound changes, origin of the item and present-day pronunciation. For example, in Figure 1, Older Scots shows a spelling variant <hemp> deriving from Old English [henep]. The Older Scots phone [hɛmp] is reconstructed via the established and known changes that affected the variety. This grapho-phonological mapping has been compiled for each spelling variant recorded in FITS, yielding a total of 110,000 phonologically parsed tokens, and an extensive spelling convention profile of each sound, visually presented by *Medusa* (Figure 2).

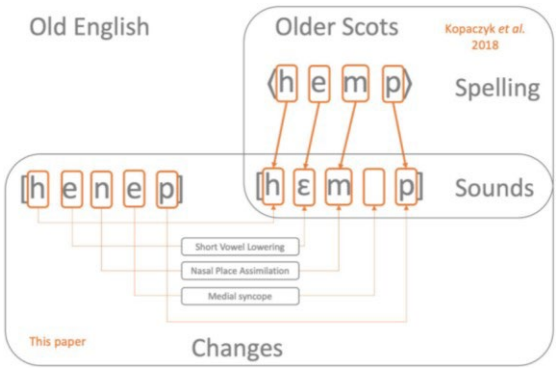

Figure 1: Grapho-phonological parsing of *hemp* in FITS in Molineaux et al. (2023: 50)

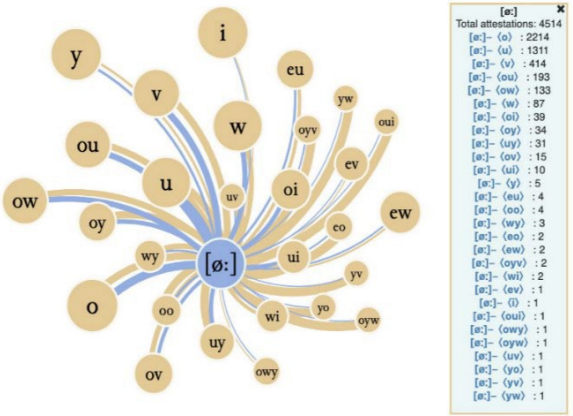

Figure 2: *Medusa* visualisation of [øː] graphemes in Molineaux et al. (2023: 61)

The model provided by FITS allows quantitative research on the development of specific phonemes across various phonotactic and grammatical contexts, including localising sound changes via spelling changes in time and space. While FITS itself is by nature specific to Scots, the methods of grapho-phonological parsing have already been used within the *Corpus of Historical Mapudungun.* This method can be applied to any pre-standard language variety (stage) that has written evidence and can also offer novel insights into the phonological developments of German varieties such as Middle High German or Early New High German.

Katharina Zeh (Universität Wien), Julia Neidhardt (TU Wien) Hannes Fellner, Andreas Baumann (Universität Wien)

Zusammenhänge zwischen Digitalisierung und linguistischer Diversität: eine Roadmap

Etwa 7.000 Sprachen werden weltweit gesprochen, von denen rund 4.000 als indigene Sprachen eingestuft sind und von etwa 5.000 ethnischen Minderheiten in 70 Ländern gesprochen werden. Trotz dieses Reichtums ist die sprachliche Vielfalt in den letzten Jahrzehnten erheblich zurückgegangen, wobei 40% der Sprachen weltweit vom Aussterben bedroht sind (Harmon & Loh, 2010). Die Gründe für diesen Rückgang sind vielfältig. Aktuelle Forschungen von Bromham et al. (2021) zeigen, dass die Gefährdung von Sprachen das Ergebnis komplexer sozioökonomischer und geografischer Faktoren ist. Ein Faktor, der in diesem Zusammenhang jedoch bisher wenig untersucht wurde, ist die Digitalisierung – ein zunehmend einflussreicher Aspekt. In diesem Vortrag stellen wir unser Projekt DIGILINGDIV vor, das darauf abzielt, Erkenntnisse aus der evolutionären Linguistik, der Informationstechnologie und der Datenwissenschaft zu kombinieren, um die Rolle der Digitalisierung für die sprachliche Vielfalt zu bewerten.

Die Auswirkungen der Digitalisierung auf die sprachliche Vielfalt sind ambivalent. Einerseits erleichtert sie die Dokumentation und den globalen Zugang zu sprachlichem Wissen. Digitale Werkzeuge haben die geisteswissenschaftliche Forschung erheblich vorangebracht und zur Bewahrung kulturellen Erbes beigetragen (Giglitto et al., 2023). Soziale Medien und andere Plattformen ermöglichen es Sprechern von Minderheitensprachen, miteinander zu kommunizieren, was ihr Überleben unterstützen könnte. Ebenso können maschinelle Übersetzungstechnologien die mehrsprachige Kommunikation fördern. Andererseits offenbart die Verteilung von Sprachen in digitalen Räumen erhebliche Ungleichheiten. Die digitale Kluft zeigt sich deutlich auf Plattformen wie Twitter („X“), wo über die Hälfte der Nachrichten in nur drei Sprachen – Englisch, Japanisch und Spanisch – verfasst wird, die jedoch nur von einem Achtel der Weltbevölkerung gesprochen werden (Pfeffer et al., 2023).

In diesem Vortrag präsentieren wir die Roadmap unseres DIGILINGDIV-Projektes. Wir beschreiben Daten und Methoden, mit denen wir auf unterschiedlichen Wegen den Zusammenhang zwischen Digitalisierung und Sprachdiversität untersuchen wollen: (i) Eine quantitative Korrelationsanalyse von Maßen und Kennzahlen zu den beiden Domänen; (ii) Die Erstellung eines globalen Textkorpus, das Sprachdiversität im digitalen Raum abbilden soll; und (iii) eine Umfrage zur Sprachwahl im digitalen und nichtdigitalen Diskurs. Dabei werden auch erste Zwischenergebnisse präsentiert.

Michelle van de Bilt (University of Vienna)

“How music changes through the years”: A sentiment analysis of Queen’s song lyrics

British rock band Queen are widely considered to be one of the most influential musical groups of the twentieth century, with a career spanning more than twenty years in the original line-up that included their late singer Freddie Mercury. It has been noted that over the course of Queen’s career, their musical style changed considerably, as is for example evident in their selection of musical instruments and the density of their arrangements (Bizzo n.d.). It seems plausible, therefore, that corresponding changes could be found in their lyrical style. To gain a first insight into the lyrical development in Queen’s music, a sentiment analysis was performed on their song lyrics.

For this purpose, I created a corpus containing lyrics to 143 Queen songs released on the band’s studio albums recorded by the original line-up (1973-1995). Four sub-corpora were created to represent the different artistic phases of Queen’s career: Period 1 from 1973 to 1976, Period 2 from 1977 to 1980, Period 3 from 1981 to 1986, and Period 4 from 1987 to 1995. The periodisation was based on both significant events in the band’s history as well as changes in musical style. As the band was more productive in their first artistic period than in any later period, the corpus was balanced so that each subcorpus consisted of 6,650 tokens. The corpus was then combined with a dataset of sentiment norms for nearly 14,000 English lemmas (Warriner, Kuperman and Brysbaert 2013). For each period, the mean arousal, dominance, and valence of Queen’s lyrics were calculated. To determine statistical significance, ANOVA-tests were performed on transformed data.

Natalia Borza (Pázmány Péter Catholic University)

Validating the BEMDI-metre, an analytical instrument that detects basic emotions in discourse

Validating the BEMDI-metre, an analytical instrument that detects basic emotions in discourse

All humans have an innate set of basic emotions (BEMs) that are clearly recognisable across cultures (Ekman, 1992, 2023; Ekman and Cordaro, 2011). Based on cross-cultural studies, emotion psychologist Ekman (1992) concluded that there is a handful of BEMs: anger, fear, enjoyment, sadness, disgust (as core emotions), and additionally contempt and surprise.

Ekman's set of BEMs is apparently the most clearly expressed and lexically labelled in language (Bahn, 2014). Nevertheless, this psychological taxonomy of emotions has not been applied in discourse analysis.

The aim of the present validation study is to investigate whether the application of Ekman's (1992) taxonomy, which has been used in psychology for decades, provides reliable results in discourse analysis. In order to reach this aim, I develop an analytical instrument that is capable of detecting BEMs in discourse, the BEMDI-metre. The BEMDI-metre consists of two parts, 1) a manual for the annotators and 2) a list of exploratory questions. The 13-page- long manual, which is based on the literature by Ekman and his team (2023) provides a common ground for the annotator, while the list of exploratory questions ensures that the particular BEM is textually present in the discourse.

To validate the operationalization of Ekman's (1992) taxonomy of BEMs in discourse analysis, a corpus of social media comments was compiled. The genre of comments was chosen on the basis that it characterised by the use of emotional language. Furthermore, the probability of identifying a range of different emotions in a collection of short texts is greater than in one single text. From a pool of about half a thousand (N=495) social media comments (for more details on the selection principles see Borza, 2023), 45 comments were selected in such a manner that every BEM appeared approximately 20 times in the corpus. The corpus was annotated independently by six annotators, who received instruction in the application of the BEMDI-metre. The annotators discussed their tagging, which led to a consensus on each comment regarding the presence of a particular BEM. Next, the percentage with which each annotator identified each BEM was calculated and it was compared to the statistical probability.

The results of the validation study indicate that the set of BEMs can be identified in discourse with the application of the BEMDI-metre. Furthermore, the tool is suitable for the detection of the presence of multiple BEMs in discourse. Thus the BEMDI-metre is an analytical instrument that can be used to operationalise Ekman's (1992) taxonomy in discourse analysis. The analysis yielded interesting results that lend themselves to further, albeit speculative interpretation. Arousal values were significantly higher in Period 3 than in both Period 1 and 2, which might be connected to Queen's tentative foray into political themes during this period. The results that were particularly noteworthy concern Period 4, in which the band members were coping with Mercury's impending death. During this time, their lyrics were significantly less dominant than in Period 1. Perhaps counterintuitively, it was found that Queen's lyrics were significantly more positive in their fourth artistic period, i.e. their final years, than in their second, at the start of which they had just risen to fame. This could reflect an increased sense of group identity and friendship among the band members during their final years together (Doherty 2011: 74, Rogan 2021).

The results show that in terms of sentiment, Queen's lyrics changed significantly over the course of the band's career. Further analysis of Queen's lyrical development could investigate the songs' linguistic complexity within the different artistic periods.

Nicola Brocca (Universität Innsbruck)

Förderung und Ressourcen für die digitalen Linguistik in Österreich am Beispiel des Ladder Web-Projekts

Das Projekt LadderWeb ist eine innovative Webanwendung, die die Annotation von Absagen und Anfragen auf pragmatischer Ebene unterstützt. Die Anwendung basiert auf manuell annotierten Korpora, darunter Ladder und DisDir, welche als Trainingsdaten für ein neuronales Netz dienen. Dieses Modell nutzt fortschrittliche Techniken der künstlichen Intelligenz, um aus den annotierten Daten präzise Annotationen abzuleiten. Die genannten Korpora stammen aus früheren Forschungsprojekten (Nuzzo & Cortés Velásquez 2022; Brocca et al. 2023) und sind in wissenschaftlichen Repositorien wie Zenodo

und ARCHE archiviert. Sowohl die Daten von LadderWeb als auch der Code der App selbst sind unter einer CC-BY-Lizenz verfügbar und folgen den FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable).
Die Entwicklung von LadderWeb schreitet dynamisch voran. Aktuell basiert das System auf zusätzlichen Trainingsdaten, die eine Genauigkeit von über 75 % bei den Annotationen ermöglichen. Detaillierte Ergebnisse dazu werden auf dem Workshop „Natürliche Sprache und Künstliche Intelligenz: Neue Perspektiven für linguistische Studien“ bei der ÖLT 2024 präsentiert. Geplante Erweiterungen der Anwendung beinhalten die Integration von Large Language Models (LLMs), welche die Funktionalität weiter steigern sollen.
In der vorliegenden Präsentation wird die Entstehung und Weiterentwicklung des LadderWeb-Projekts im Kontext verschiedener Förderungen und Infrastrukturen beleuchtet. Dies umfasst unter anderem die 5. und 8. Ausschreibung des Förderprogramms
„Digitalisierung und Informationsextraktion für die Digital Humanities“ (DI4DH) des Forschungszentrums Digital Humanities der Universität Innsbruck in den Jahren 2020 und 2021 sowie die Unterstützung durch das FZDH-Netzwerk. Darüber hinaus war das CLARIAH-AT-Förderprogramm von 2022 entscheidend für die Weiterentwicklung des
Projekts. Die digitale Infrastruktur von ARCHE, einem CLARIN B-Zentrum, und die Vernetzung über das CLARIN Café sowie Veranstaltungen wie die Roadshow der Summerschool 2024 haben ebenfalls maßgeblich zur Verbreitung der App beigetragen.
Diese Präsentation zielt darauf ab, eine Diskussion über aktuelle Förder- und Vernetzungsmöglichkeiten für digitale linguistische Projekte in Österreich anzuregen. Ziel ist es, zukünftige Chancen aufzuzeigen und den Austausch zwischen Forschenden, Institutionen und Fördergebern zu fördern, um die digitale Linguistik in Österreich weiter voranzutreiben.

Elisabeth Gruber-Tokić, Milena Peralta Friedburg (Universität Innsbruck)

Item Plasia von Czïmer Nicleins weib an der port czinst ii guntsch wein[1]: Digital Processing and Analysis Late Medieval Socio-Economic Networks using CIDOC-CRM

Princely or noble estate registers (Urbaria) represent official records of the ownership of farms and estates and were written in the Middle Ages until Early Modern times. These documents contain information regarding the claimed levies and services by the landlords. As a result, estate registers give us an insight into manorial structures and agricultural resources as well as social relationships. The precise listing of the names of estates and the individuals liable to pay the tax makes estate registers a valuable source for Onomastics (anthroponyms and toponyms) and History although the data might be imbalanced regarding late medieval society.
The interdisciplinary research project ViTA „Vernetzung im Tiroler Alpenraum“[2] (2023 –2025) is funded by the programme go!digital 3.0 of the Austrian Academy of Sciences and conducted at the University of Innsbruck. The project aims to make a first step towards an interactive map revealing the relationships between estates, associated levies, persons and the regional distribution of produced goods.
This map is based on two late medieval estate registers of the historical County of Tyrol: 1) TLA, Urb. 74/3 relates to the tyrolean noble family Starkenberg and dates to approx. 1390. 2) TLA, Urb. 1/2 (1406-1412) contains information on the property of Duke Frederic IV of Habsburg.
The talk will put emphasis on the research objectives and the methodology for data processing and semantic representation in the research project ViTA. We discuss challenges of the historic-semantic annotation in general and of Early New High German anthroponyms in particular, e.g., identification of women and aspects of historic multilingualism.
In addition, as estate registers reflect the socio-economic networks of the time, we raise the question, if these might be semantically represented using the ontology CIDOC-CRM and further visualised and analysed with a Knowledge Graph. For this attempt the research project ViTA implements the extensions CRMsoc – which currently exists as a draft – as well as CRMtex to determine their viability regarding the presented historical documents.

[1] TLA, Urb 1/2: 211v.
[2] Manorial (Economic) Networks in the Medieval Tyrol: Mapping and Visualisation.

Urbare (< mhd. *erbërn* < ahd. *irberan* „etwas erzeugen; gebären“) als mittelalterliche und frühneuzeitliche Verwaltungsdokumente sind Verzeichnisse von Besitzrechten und geforderten Abgaben einer Grundherrschaft und erlauben uns Einblicke in grundherrschaftliche Strukturen sowie landwirtschaftliche Ressourcen. Die präzise Auflistung von Eigentum, Abgaben und den jeweiligen steuerpflichtigen Personen machen Urbare – trotz des sozialen Ungleichgewichts – zu wertvollen Quellen für die historische Onomastik sowie Geschichtswissenschaften.
Das interdisziplinäre Innsbrucker DH-Projekt ViTA – „Vernetzung im Tiroler Alpenraum“[3] (2023–2025), gefördert durch ÖAW go!digital 3.0., wählte zwei spätmittelalterliche Urbare des Tiroler Landesarchives (TLA) zu Innsbruck für die Digitalisierung und semantische Aufbereitung der enthaltenen Daten aus, um diese in einer interaktiven Karte zur Verfügung zu stellen. Die bisher unveröffentlichten Quellen TLA, Urb. 74/3 (ca. 1390) und TLA, Urb. 1/2 (ca. 1406-1412) enthalten Aufzeichnungen zu den Besitztümern und zugehörigen Abgaben der Tiroler Adelsfamilie Starkenberg und Herzog Friedrich IV. von Habsburg in der historischen Grafschaft Tirol.
Schwerpunkte des geplanten Beitrags sind die Vorstellung des Forschungsprojektes sowie der Methodologie der Datenaufbereitung und semantischen Datenmodellierung. Dabei diskutieren wir Erkenntnisse und Herausforderungen der historisch-semantischen Annotation im Allgemeinen und vor allem im Bezug auf frühneuhochdeutsche Personenamen. Darunter fallen beispielsweise die Identifikation von Frauen innerhalb von Verwaltungsdokumenten, die Grenzen von Personennamen, sowie Aspekte historischer Mehrsprachigkeit.
Außerdem beschäftigen wir uns mit der Frage, ob und inwiefern historische sozioökonomische Netzwerke und Personennamen unter Nutzung der Ontologie CIDOC-CRM und deren *extentions* (z.B. CRMsci, CRMinf) modelliert und in einem Knowledge Graphen visualisiert und analysiert werden können. Durch den Einsatz der Erweiterungen CRMsoc – die bisher nur als Draft existiert – sowie CRMtex erforscht das Projekt ViTA gleichsam auch deren Viabilität für das bearbeitete historische Quellmaterial

Michael Gassner, Christina Katsikadeli, Thomas Klampfl (Austrian Academy of Sciences)

Etymological and sociolinguistic information extraction from digital historical dictionaries (according to the DLGenR and DLYS projects)

Based on recent investigated data from two FWF-projects on language contact between Greek and Herbrew/Aramaic (Digital Dictionary of Loanwords in the Midrash Genesis Rabbah, *DLGenR*, and the follow-up project: Loanwords in Yalkut Shimoni, DLYS), we present case studies in which the investigation of loanwords deals with complex contact situations between these two languages as well as with a large time span of interaction, which again necessitates the diachronic, diatopic as well as the diaphasic investigation and exploitation of these linguistic data in their own right. Both digital dictionaries cover text samples from the whole diachrony of Ancient and Early Medieval Judaism (5th to the 12th century CE) and will comprise almost 2/3 of the Greek loanwords in Post Biblical Hebrew/Aramaic (from the earliest Mishnaic to the latest Rabbinic sources) in a digital format.
Although several Greek loanwords in the Jewish literary tradition might seem straightforward, the bulk of these lexical items involve the examination of whether this specific word is Greek, and if this is the case, which one of several possibilities could be the most –scientifically speaking– appropriate, not only in accordance with the semantics of the context, but also on the basis of other relevant linguistic criteria. Thus, for every single entry, one has to start anew the application of an entire range of phonological and morphological possibilities alongside semantics and sociolinguistic criteria. Consequently, the dictionary type of DLGenR and DLYS shifts from a lexicon that initially focuses on the word history (*histoire des mots*) of specific loanwords to an etymological one, which explores the history of origin (*histoire des origins*). In order to capture the range of these various factors, all entries follow a set of etymological criteria, according to the “maximal” model proposed by Hoffmann / Tichy (1980) (adopted to our needs):

I. Occurrence (reality; chronology; region; register and frequency; variation);
II. Written attestation: (genre; manuscripts);

[3] Manorial (Economic) Networks in the Medieval Tyrol: Mapping and Visualisation.

III. Linguistic authenticity ("genuine" loanword vs. ad-hoc formation);
IV. Tracing the word meaning (philological/semantic interpretation; thematic environment; us age; semantic shift);
V. Tentative reconstruction (phonological adaptation; morphological adaptation; compounding);
VI. Etymological Linkage (motivation of word formations; semantic interpretation; proposal ranking)

In addition, for the first time in the relevant research fields, we also draw attention to the indispensable etymological criterion of relative chronology, which has been impossible to follow until now due to the absence of chronological classification and the disregard of morpho(phono)logical constraints and sociolinguistic factors in the treatment of these loanwords.
Compared to print editions, we show how the digital encoding in XML/TEI-LexO already used in both projects makes it possible to easily merge the entries into a single web-client and align the findings with existing projects and research on the Ancient and Medieval Mediterraneanas a sociolinguistic area – beyond the study of Greek and Aramaic.

Veronika Engler (Austrian Academy of Sciences)

The WIBARAB Feature Database: Developing a Data Model for a TEI-Based Linguistic Database

A central part of the ERC funded WIBARAB (What Is Bedouin-type Arabic?) project is its TEI-based database which has been constructed to facilitate linguistic analysis of a great number of Arabic varieties. The WIBARAB database currently includes approximately 100 linguistic features, mainly from phonology, morphology, and syntax, but also a selected set of lexical and phraseological items. The finished database will contain linguistic data from approximately 300 publications, fieldwork campaigns and personal communications with experts about around 320 Arabic varieties. Many of the investigated varieties are spoken by (formerly) nomadic communities which is due to the project's research focus on the Bedouin-sedentary split of spoken Arabic. The WIBARAB database is not only the first all-Arabic database of this size, but it is also the first one to focus on the varieties spoken by Bedouins.
The general aim of the WIBARAB database is to facilitate and enhance linguistic analysis and to make a wide range of data easily accessible and quickly comparable to scholars and anyone interested in variation within spoken Arabic. In order to achieve this, we needed to develop a flexible and easily customisable data model, thus choosing a TEI-based approach. Generally, in the development of the data model we encountered three main challenges: (1) the overall wish to remain faithful to the data and minimizing the influence of authors' and/or researchers' interpretations, (2) the complex relation between the spoken varieties, locations and the speech communities, and (3) the diversity of the sources and the resulting diversity of the data.
In this presentation, I will discuss these three main challenges and how they essentially shaped our data model which provides the 'skeleton' for the WIBARAB database that is flexible and dynamic enough to be customised in order to faithfully reflect the complexity of the methodological and (socio)linguistic variation we found in the sources.

Saniya Irfan (IIT Delhi)

Navigating Linguistic and Technical Challenges in encoding Urdu Marsiya: Developing a Digital Scholarly Edition with TEI

Despite its popularity and usage over the past 25 years, TEI has rarely been applied to South Asian languages, particularly in Urdu, which contains a significant amount of literature in its original manuscript format. These manuscripts are circulated as scanned, un-editable, and static PDFs. Despite its rich literary tradition, Urdu is facing a decline in accessibility and readership, therefore, there is a pressing need for digital editions of Urdu literature. The problem with ancient and pre-modern Urdu literary texts is in their utilisation of distinctive calligraphic techniques, often lacking any references in their first prints. The lack of digitization of old Urdu literature, with its unique and unfamiliar calligraphic forms and absence of metadata, is a challenge for young Urdu language users to fully appreciate and understand the nuances within the texts. A digital archive of Urdu literature can significantly mitigate the issues by providing accessible, interactive, and enriched textual experiences. Text encoding with

TEI standards can include comprehensive metadata, contextual annotations, and translations which help overcome the barriers posed by traditional calligraphic forms. This segment of my research work outlines the development of the first TEI-encoded digital archive of Urdu Marsiya (elegies), focusing on a significant Urdu corpus from 19th-century Lucknow. These elegies, commemorating the Battle of Karbala and the martyrdom of Prophet Muhammad's grandson, are pivotal to Urdu literary heritage. The approach utilizes an Optical Character Recognition (OCR) engine to transcribe these texts from their original Perso-Arabic script. This transcription process involves significant manual correction to accommodate the unique calligraphic styles of Urdu, which often omit essential diacritical marks and employ context-specific letterforms. Following transcription, the texts are encoded according to TEI standards, facilitating detailed linguistic and stylistic analysis. The presentation will discuss the technical challenges encountered such as issues with OCR accuracy due to the script's right to-left orientation of the Perso-Arabic script, the inadequate word separation in Urdu and complex calligraphy. I outline the technical solution implemented, which includes a programmatic way of (i) automating text extraction from PDFs using OCR technology (ii) extracting relevant entities such as people, places, events, concepts used in the text using Named Entities Recognition techniques in Natural Language Processing (NLP) and finally, (iii) transforming the extracted text into a TEI encoded Digital Scholarly Edition. The broader implications of this work for the field of Digital Humanities in South Asia and the preservation of regional languages are also considered, highlighting the project's contribution to global scholarship and the promotion of Urdu literary studies.