Workshop on
Computational Linguistics providing Paths to Transdisciplinary Research

Dedicated to Ulrich Heid.

Contents

1

Preface

Wenn man bedenkt, dass die ersten Versuche, computationell Sprachen zu verarbeiten (im Kontext der maschinellen Übersetzung), nun bereits über siebzig Jahre zurückliegen, ist es unserer Ansicht nach an der Zeit, zumindest Teilaspekte der Sprachtechnologie und ihren transdisziplinären Anspruch in einem Workshop zu beleuchten. Zwischen Linguistik und Informatik liegend, hat die Sprachtechnologie beide Disziplinen in den vergangenen Jahrzehnten signifikant beeinflusst und man kann behaupten, dass sich Linguistik und Informatik über ihre transdisziplinäre Brücke gegenseitig bereichert haben.

Die Sprachtechnologie beschränkt sich schon lange nicht mehr auf ihre traditionellen vier ‚Räder' Morphologie, Syntax, Semantik und Pragmatik, mit denen sie Texte für linguistische Zwecke analysiert und generiert. Durch Einfluss der Informatik sind statistische Werkzeuge und – heutzutage – neuronale Netzwerke in den selbstverständlichen Gebrauch und in die Curricula der dazu angebotenen Studiengänge eingegangen.

In den letzten Dekaden hat sich darüber hinaus das transdisziplinäre Angebot der Sprachtechnologie auf viele weitere Fächer ausgedehnt, in denen ebenfalls mit Texten gearbeitet wird. Nicht nur spielt die Sprachtechnologie schon länger eine wichtige Rolle für die (Fach-)Lexikographie, zum Beispiel bei der Identifizierung lexikalischer Bedeutungen in Textsammlungen. Mittlerweile erschließt sie im Bereich Digital Humanities traditionelle Fächer wie die Literatur-, Sozial-, Geschichts- und Politikwissenschaft.

Ohne das Vorwärtsstreben universitärer Forschung wäre die Sprachtechnologie heute sicher eine nicht so breit gefächerte Disziplin. Gerade hier hat Ulrich Heid einiges bewegt und erreicht. Bei dem am 17. Oktober 2019 stattfindenden Workshop *Computational Linguistics providing Paths to Transdisciplinary Research* zum Anlass seines sechzigsten Geburtstags betrachten wir – bezogen auf seine Forschung – die Sprachtechnologie aus verschiedenen Perspektiven.

Wir beginnen mit dem Beitrag *Reusability of Lexical Resources in Eurotra.* Hier beschreiben Pius ten Hacken und Folker Caroli die Wichtigkeit lexikalischer Ressourcen für die maschinelle Übersetzung in den achtziger Jahren. Daran schließt Kurt Eberle an mit seinem Beitrag *Balancing Corpora for Machine Translation*, bezogen auf das SFB-732-Projekt *Incremental Specification in Context* aus den 2000er Jahren. Kerstin Jung (geb. Eckart) berichtet im Anschluss ebenfalls über dieses Projekt, nun aus der erst seit wenigen Jahren berücksichtigten, und doch hochrelevanten Perspektive der Prozessdokumentation, die sie in ihrer Dissertation, betreut von Herrn Heid, untersucht hat. Jung beleuchtet die Arbeit an den entsprechenden ISO-Normen, ohne deren Daten- und Schnittstellenbeschreibungen eine inter- bzw. transdisziplinäre Prozessierung von Daten kaum möglich wäre.

Mit dem Vortrag von Rufus Gouws wechseln wir erneut die Perspektive und betrachten nun den Bereich der Kollokationen, zu denen Herr Heid seit vielen Jahren arbeitet. In der Lexikographie ist das Wissen um Kollokationen essentiell und ohne Sprachtechnologie (bzw. Korpuslinguistik) nur begrenzt untersuchbar. Für die Verarbeitung der afrikanischen Sprachen Südafrikas ist der Name Ulrich Heid ebenfalls nicht wegzudenken: Theo Bothma berichtet gemeinsam mit Danie J. Prinsloo hierzu über verschiedene

Projekte und sprachtechnologische Software, die in Zusammenarbeit mit ihm entstanden sind.

Während der Vortrag von Bothma und Prinsloo sich auf lexikographische Projekte zu südafrikanischen Sprachen konzentriert, werden Sonja Bosch und Gertrud Faaß über weitere sprachtechnologische Ressourcen und Werkzeuge sprechen, von denen viele in der letzten Dekade nicht ohne Herrn Heids Beteiligung und seinen Rat entstanden wären. Auch bezüglich Fachsprachen war und ist Herr Heid sehr aktiv: Michael Dorna von der Robert Bosch GmbH und Anna Hätty von der Universität Stuttgart berichten in ihrem Vortrag *T&O - Terminology Extraction and Ontology Development* über ein terminologisches Kooperationsprojekt unter der Ägide von Herrn Heid. Im Anschluss daran beleuchtet der Vortrag *Specialised Lexicography and Term Variation* von Theresa Kruse und Laura Giacomini die bisherige Hildesheimer Zeit mit der Beschreibung von Qualifikationsarbeiten zu Fachlexikographie und Termvariation in den Gebieten Mathematik und Technik.

Auf den transdisziplinären Anteil der Sprachtechnologie an den Digital Humanities geht Gerhard Lauer ein: In *Social Reading and Writing Media* fasst er zusammen, wie Beiträge aus sozialen Medien in Echtzeit-Analysen untersucht werden können. Johannes Schäfer, Anna Moskvina und Fritz Kliche bleiben im Bereich Digital Humanities. Sie stellen vor, wie für bibliographische Zwecke mit

sprachtechnologischen Methoden die Inhalte wissenschaft-
licher Publikationen erfasst und verschlagwortet werden
sollen; eine andere Arbeit befasst sich mit der Erkennung
von unangemessenen Inhalten wie Hate Speech in sozialen
Medien.

Die Mitarbeiterinnen und Mitarbeiter der Computerlinguistik-
Gruppe des Instituts für Informationswissenschaft und Sprach-
technologie an der Universität Hildesheim: Gertrud Faaß, Laura
Giacomini, Max Kisselew, Fritz Kliche, Theresa Kruse, Anna
Moskvina und Johannes Schäfer

Juni 2019

2

Program

10:00	Arrival, Registration with coffee on offer
11:00	Welcome: Wolfgang-Uwe Friedrich President University of Hildesheim
11:20	**Session 1: European Machine Translation Project(s)** Welcome: Folker Caroli (Hildesheim)
11:25-11:45	Pius ten Hacken (Innsbruck) and Folker Caroli (Hildesheim)
11:45-11:50	Session 1 round-up (Pius ten Hacken)
11:50-11:55	**Session 2: Working on SFB 732 "Incremental Specification in Context" at University of Stuttgart** Welcome: Kurt Eberle (Heidelberg)
11:55-12:15	Kurt Eberle (Heidelberg)
12:15-12:35	Kerstin Jung (Stuttgart)
12:35-12:40	Session 2 round-up (Kerstin Jung)
12:40-12:45	**Session 3a: Advances in Theoretical Lexicography** Welcome: Theo Bothma (Pretoria, South Africa)
12:45-13:05	Rufus Gouws (Stellenbosch University, South Africa)
13:05	Lunch Break (buffet)
14:15	**Session 3b: Lexicographic Projects & African Languages Issues**
14:15-14:35	Theo Bothma (with support of Danie J. Prinsloo, Pretoria, South Africa)
14:35-14:55	Sonja Bosch (UNISA, Pretoria, South Africa) and Gertrud Faaß (Hildesheim)
14:55-15:00	Session 3 round-up (Rufus Gouws)
15:00-15:05	**Session 4: Terminology and Terminological Projects** Welcome: Laura Giacomini (Hildesheim)
15:05-15:25	Michael Dorna (Bosch CR, Renningen) and Anna Hätty (Stuttgart)
15:25-15:45	Laura Giacomini and Theresa Kruse (Hildesheim)
15:45-15:50	Session 4 round-up (Laura Giacomini)
15:50-16:05	Coffee Break
16:05-16:10	**Session 5: Digital Humanities** Welcome: Caroline Sporleder (Göttingen)
16:10-16:30	Gerhard Lauer (Basel)
16:30-16:50	Anna Moskvina, Johannes Schäfer, Fritz Kliche (Hildesheim)
16:50-16:55	Session 5 round-up (Caroline Sporleder)
16:55-17:15	**Round up: Final Discussion and Summary** (Ulrich Heid)
17:15	End of the workshop
17:15-18:30	University Reception: Faculty 3 and guests
18:45	Dinner in local restaurant: Os - Das Marktrestaurant (Markt 7, 31134 Hildesheim)

3

Reusability of Lexical Resources in Eurotra

– Pius ten Hacken, *Universität Innsbruck*, Folker Caroli, *Universität Hildesheim*

This talk focuses on the early period of research activities by Uli Heid. It describes the main context of natural language processing research in Europe at this time, the Eurotra Project. Eurotra was the machine translation (MT) project of the European Community in the 1980s. It was a project that was unique in its design and size, aiming to develop a truly multilingual transfer system for the nine EC languages at the time. Its implementation and organization reflected an approach to MT that was mainly interested in theoretical questions of linguistic

modelling. Most research was devoted to syntactic and semantic issues. Only towards the end of the project, the so-called lexical bottleneck became a prominent issue. It was discovered that for a proper operation, the system would need large monolingual and bilingual dictionaries, whose development would require a significant resource. The Dictionary Task Force (DTF) was set up to plan and monitor this work.

Thus, at the end of the main Eurotra project in 1990, one of the follow-up projects was a tender named "Reusability of Lexical Resources". The main contractor for this tender was Ulrich Heid. The objective of this project was to produce an overview of the state of the art in the reusability of lexical resources. It consisted of two components, one producing an inventory of existing digital lexicographic resources and one aiming to set up a standard for the encoding of lexical information. The outcomes of the project lead to two conceptions to building up dictionaries for natural language processing systems, on one hand the extraction of lexical information from existing machine-readable resources for human use, on the other the production of lexical resources by computer-based exploitation of large text corpora. This second orientation has become one of the main research fields of Uli Heid.

4

Balancing Corpora for Machine Translation

– Kurt Eberle, *University of Heidelberg*

Text analysis and translation have made big progress over the last couple of years by adding the conception of neural networks to the paradigms of Machine Learning. Text classification could be significantly improved as well as Machine Translation. However, a prerequisite is that very large corpora are available for training which is not the case for any language or language pair or domain. What is even worse is the fact that large data not necessarily prevent learning from obtaining biased results. In Machine Translation, for instance, it could be shown that good neural systems that have been created from large data,

though producing remarkably fluent and linguistically rich results, nevertheless show many errors in adequacy for texts with high percentage of technical terms and unusual selection from the range of readings of words, phrases and sentences. To remedy on this it is helpful to train systems on corpora that are not only large but balanced enough to represent an optimum of the use cases one is interested in and that avoid unwanted biasing towards specific meanings and translations.

In the special research area 732 on *Incremental Specification in Context* and in particular in its project on linguistic tools, guided by Uli Heid, we had developed devices for extracting and labeling references for various phenomena from text. Optionally, the tools allowed to integrate syntactic and semantic analyses of the texts and the use of semantic and pragmatic predictions as deducible from discourse representation theory (DRT) and its model theory in order to obtain highly filtered and reliably labeled data.

These devices have been continuously developed further. The talk reports about the current integration of knowledge from distributional semantics into the extraction and labeling tools for the purpose of obtaining better information about the distribution of readings and for compiling more balanced corpora as a basis for classifiers. In the approach we report on, the results of the classifiers are evaluated by the labeling tools in turn, such that the quality of both, extraction and labeling tool and classifier are optimized in a bootstrapping approach.

5

Representing underspecification – DIN and a database

– Kerstin Jung, *University of Stuttgart*

Inspecting lexical ambiguities, such as the ambiguity between object, event and state readings of German *-ung*-nominalizations, includes the inspection of many examples. In project B3 of SFB 732, next to corpus-based examples, also a set of analyses from several tools had to be explored and stored.

Since most linguistic representation tools are based on completely specified analyses and do not provide means to natively

store and compare several analyses from the one layer (e.g. syntax) for the same example, the relational database B3DB emerged from the project. Starting from a detailed set of tables for texts, paragraphs and sentences, the database developed into a generic structure with two layers: a macro layer for process description and a micro layer for a graph representation of analyses, including their degree of (under)specification. The design of the micro layer in turn is closely connected to the development of the annotation representation standards by ISO's TC 37/SC 4, the sub committee *Language resource management* of the technical committee *Language and terminology*. This sub committee is mirrored by the *Arbeitsausschuss Sprachressourcen* from DIN's *Normenausschuss Terminologie*. With its roots both in application and standardization, the B3DB is one of the outcomes of project SFB 732 B3 and with its focus on joint representation, process metadata and comparison exploration it has been applied for further projects such as the combination of information status and prosody.

6

Collocations and dictionaries – then, now and in future

– Rufus Gouws, *Stellenbosch University*

The past four decades have witnessed an active discussion of the presentation and treatment of collocations in dictionaries. These discussions have been characterized by significant differences, a number of suggested changes and improvements, a focus on dictionaries in the transition from the printed to the online medium and a variety of approaches regarding the status of collocations as items in dictionary articles. A constant factor

in these deliberations has been the name of one of the major discussion partners, namely Ulrich Heid.

The appropriate choice of collocations often distinguishes mother-tongue speakers from those speakers using a second or third language. As information tools directed at the real needs of real users dictionaries need to take cognizance of problems regarding the choice and proper use of collocations. This applies to both monolingual and bilingual dictionaries and to dictionaries dealing with language for general purposes but also those directed at languages for special purposes. Attempts to ensure an optimal treatment of collocations in dictionaries remain a real challenge that demands innovative commitment on the level of both metalexicography and the lexicographic practice. The planning of the treatment, however, should be preceded by a planning of the selection and presentation of collocations. In his in depth research over many years Ulrich Heid has committed himself to finding solutions that can meet this challenge. His work with regard to the linguistic nature of collocations and their treatment in printed dictionaries was complemented by research regarding the possibilities that the transition to language technology within the framework of digital humanities could offer the (meta)lexicographer.

Collocations are at the center of this talk. The talk takes a contemplative look at the development of (meta)lexicographic approaches to collocations and the role Ulrich Heid has played in this regard. The contemplative look is supplemented by a transformative look that focuses on future possibilities and gives a re-appraisal of the status of collocations as items in dictionary articles. The need to elevate collocations from mere microstruc-

tural items addressed at the lemma sign to treatment units that function as address of a variety of item types is discussed. With reference to the online environment attention is drawn to the accommodation of collocations in specific search zones and the introduction of a rapid access structure that can guide a user directly to collocations without using the lemma sign of the relevant article as guiding element.

7

Lexicographic Projects

– Theo Bothma, *University of Pretoria*

The paper will give a brief overview of various end-user guidance devices within the framework of the Function Theory of Lexicography for text production, text reception and the cognitive function, at the hand of the taxonomy the authors developed in conjunction with Ulrich Heid, based primarily on their earlier collaboration. Two devices for text production and text reception will be discussed in more detail, viz.:

1. For text production

    (a) The Sepedi copulative decision tree

(b) The Sepedi writing assistant

The text production tools can be made accessible to end-users in different ways:

(a) Incorporated into an e-dictionary
(b) Incorporated into a word processor
(c) As a stand-alone tool

The Sepedi writing assistant makes use of data from the dictionary database, and the dictionary as such is not consulted by the end-user; the dictionary is therefore 'hidden' from the end-user. Such a database requires specific data and a specific structure; some issues around this will be discussed. End-users that would like more detailed information about any step in the above examples, have, at each node, the option to drill down to more detailed explanations.

2. For text reception

(a) Amazon Kindle – linking texts to e-dictionaries and other information sources
(b) Texts linked to e-dictionaries and/or Google via the operating systems or the browser, e.g. iOS, Android, and Chrome/Firefox on Windows

For text reception, the e-dictionary does not necessarily contain all the information the end-user may require to satisfy their information need (or an expanded – new

– information need), and the end-user may require access to additional information sources, e.g., Wikipedia, Google, and other Google tools.

To access more information (for the specific or expanded information need, to learn more about the specific aspect, or even simply because of inquisitiveness), end-users have the option to drill-down, to filter information, to link to additional sources etc. This expanded access fulfils the role of the cognitive function, i.e., the end-user can learn more about the topic at hand. All the additional information is available on demand and the end-user is never exposed to an overload of information. The e-dictionary then becomes an information tool on a continuum of information tools. Examples will be provided of how such tools can be integrated in the information seeking process.

We will then argue that these developments may have an effect on dictionary usage, as well as on e-lexicography in general, specifically from an end-user perspective, and may also have an effect on dictionary construction.

8

NLP Resources and Applications for the Southern African Languages: Zulu and Northern Sotho

– Sonja E. Bosch, *University of South Africa*, Gertrud Faaß, *University of Hildesheim*

While Natural Language Processing (NLP) in the so-called developed world moves towards artificial intelligence solutions like, e.g. neural networks utilizing billions of tokens of language samples implementing chat bots or other systems that are able to communicate with humans, NLP for the Southern African languages is significantly lagging behind. One of the reasons for this is that in Southern African countries' universities do not normally offer curricula for computational linguistics studies, hence there are relatively few researchers involved in the development of NLP applications. Looking at Europe, the NLP-research community does not seem very interested in relatively small African languages, while the European linguists and African studies' community seems to have missed the Digital Humanities boat completely. Progress was made when an HLT Unit responsible for driving a new HLT strategy which included support of the research community with funds for collecting reusable resources and for developing appropriate NLP applications was established by the South African Department of Arts and Culture in 2006.

Some researchers in South Africa, however, had been collecting language data over a period of time. For instance, Danie J. Prinsloo (University of Pretoria, UP) began compiling inter alia Northern Sotho and Zulu corpora of the Southern African languages in the 1990s. One of the first NLP applications was developed by Sonja Bosch and Laurette Pretorius (both University of South Africa, UNISA) from the year 2000 on resulting in an FST-morphological analyser for Zulu.

In 2005 an opportunity arose to combine the expertise of European computational linguists and South African Bantu lin-

guists when Danie J. Prinsloo started an initiative to establish a research group together with Ulrich Heid in order to develop a Northern Sotho tagger. Subsequently Ulrich Heid added Gertrud Faaß to the team. From 2006 on, Sonja Bosch (UNISA) and Elsabé Taljard (UP) joined in and since then, not only a tagger, but a number of applications have been collaboratively developed and relatively big corpora have been collected, thus laying a solid foundation for further NLP.

This talk summarizes corpora and NLP applications for the two Bantu languages Zulu and Northern Sotho that are freely available for the research community today. For a better overview, we will include resources produced by others, such as Uwe Quasthoff and his team at the University of Leipzig and the SADiLAR (South African Centre for Digital Language Resources) community in South Africa. We will also introduce our current work. The talk serves as an invitation to the NLP community to join in working towards NLP implementations for these Bantu languages.

9

DIY beyond hammer and nails: The T&O project and its impacts

– Michael Dorna, *Bosch*,
Anna Hätty, *University of Stuttgart*

We report on the achievements and follow-up activities of the cooperation project "T&O - Terminology Extraction and Ontology Development". The project between the Institute of Natural Language Processing (IMS), University of Stuttgart,

and Robert Bosch GmbH, Corporate Research, was led by Ulrich Heid during the years 2015 to 2017. The objective of the project was to automatically extract and structure terms in the DIY ('do-it-yourself') domain. The challenges started with a heterogeneous text basis with diverse registers and text sources (e.g. encyclopedias, handbooks, marketing texts, manuals and user-written instructions). The goal was to support different people's information needs and we were faced with a different understanding and knowledge, e.g. of a DIY expert or a lay person. As a consequence, we aimed for a broad characterization of terminology, and the creation of an ontology as means to structure the non-standardized terminology. We developed a term extraction pipeline, starting from data preprocessing steps (normalization, automated linguistic annotation, lemma correction), leading to term candidate pattern search and term ranking with termhood measures. Compound splitting, term variant detection, co-reference resolution and relation extraction were further applied as basis for the ontology construction. We finally evaluated the term extractor's performance on annotated gold standard data. The outcome of the project were a range of publications and lead to a dissertation in the field. The follow-up activities explored related research questions: Amongst others, an intuitive lay understanding of terminology was empirically inquired and models for the automatic detection of terminological meaning shifts were proposed. We extended the problem to several other domains, such as cooking, hunting, chess or automotive.

10

Specialised Lexicography and Term Variation

– Laura Giacomini, *University of Hildesheim and University of Heidelberg*,
Theresa Kruse, *University of Hildesheim*

In our talk we introduce two lexicographic projects dealing with terminology variation in specialised language. The two projects explore the way in which synonymous variants belonging to the languages of technology and of mathematics can be

formally classified, annotated, extracted from corpora, and presented in LSP e-dictionaries.

The first project, carried out by Laura Giacomini, focuses on an ontology- and frame-based approach to terminology, which provides deep insights into salient semantic features of technical language, enables detailed corpus analysis and processing, and supports term variant extraction. At the core of the project is also a novel variation typology, accounting for orthographical, morphological, and syntactic alternations. A range of methods for automatically extracting different variant types from specialised corpora has been developed over the last years in close cooperation with the computational linguistics team led by Ulrich Heid at Hildesheim University. Term variation is, of course, one of the key microstructural items of the technical e-dictionary modelled in the framework of this project and intended for semi-expert users.

The aim of the second project, by Theresa Kruse, is to create an electronic dictionary for the mathematical domain of graph theory in cooperation with the Institute for Mathematics and Applied Informatics at Hildesheim University. The idea is to apply a pattern-based approach in which the structure of mathematical definitions is used to extract the terminology and to directly build an electronic dictionary with a related ontology. This dictionary shall be used by students to improve their usage of terminology.

11

Goodread Goethe. Exploring social reading and writing media

– Gerhard Lauer, *Universität Basel*

On social platforms like „Wattpad" or „Goodreads" millions of mostly young people share their stories, stories which are fan fiction, teen fiction as well as classics, surrounded by many commentaries. While in the public a lot of people fear the young generation read less and lost the ability for deep reading, a closer

look into the reading world on social platforms draws a different picture. My talk introduces into the possibilities of real time data research in reading science. New kind of big data are to handle, computer-based methods had to scale up and adapted to the new research area, and theories must be developed to understand the many data the new research encounters.

12

Digital Humanities

– Anna Moskvina, Johannes Schäfer, Fritz Kliche *University of Hildesheim*

This talk presents current research projects at the University of Hildesheim where various methods for inspecting the content of text data are developed. We apply our systems for term extraction, topic modelling and offensive language detection to a specialized corpus.

We gather both the text content and bibliographic data from digital scientific publications. Therefrom, we select a subcorpus consisting of the publications with Ulrich Heid as author or co-author. We applied a computational linguistic toolchain to analyze this data set by means of (a) term extraction by comparing the author-specific corpus to a general language cor-

pus and applying a ranking by statistical methods. In a second
part we discuss our approach of (b) topic modeling to analyze
the fields of research in Heid's publications and attempt to gain
insight in their status of transdisciplinarity in comparison to
rather isolated areas of research. Furthermore, we apply a third
system to this corpus which is also currently being developed at
the University of Hildesheim and meant to (c) detect offensive
language in social media posts. We discuss process metadata re-
quired for the creation of this corpus (and analysis) and present
an overview of results of our three systems.

Danksagung

Impressum