Abstracts KWT meeting June 18-20

Thursday Talks

Kyle Richardson (D2)
Learning for weakly-supervised semantic parsing

Recent work on Semantic Parsing has focused on learning the translation from language to formal meaning representations using parallel data, or collections of raw text-meaning pairs. While many benchmark datasets use hand-annotated meaning representations, there has been a push towards using weaker forms of supervision, or supervision that closely approximates the ground meaning or denotation of text and that cut out the need for expensive hand-engineering (e.g. Artzi, Zettlemoyer, 2011, Berant, et al. 2013). Despite the impressive results achieved by such studies, the target representations being learned are often crude, and encode very little knowledge, making it hard to learn interesting semantic generalizations and world knowledge. We discuss our efforts in using new forms of weak-supervision (e.g. textual inference judgements, plausibility judgements, grounded action) to learn refined representations for semantic parsing models, which can facilitate inference and reasoning.

Anders Björkelund (D8)
Latent structures from spurious ambiguities in transition-based dependency parsing

Transition systems for dependency parsing exhibit spurious ambiguities. We take a novel approach to training data-driven transition-based dependency parsers by letting the machine learning algorithm exploit these ambiguities in a latent fashion during training.

Agnieszka Falénska (D8)
Methods for improving dependency parsing in a cross-domain setting

Supertagging was recently proposed to provide syntactic features for statistical dependency parsing. During this talk I will describe a broad range of controlled experiments comparing this application of supertagging with another method for providing syntactic features, namely stacking. I will show how both of these methods can improve parsing of out-of-domain data.

Patrick Ziering (D11)
Cross-lingually supervised methods for parsing noun compounds

Noun compounding is a very productive type of word formation. As a result, we find many examples of this phenomenon, but very few individual tokens. Annotated data that serves compound analysis is even sparser. This has the consequence that statistical methods for analyzing noun compounds run into problems of data sparsity. In my PhD work, to circumvent the need for annotated data, I exploit cross-lingual diversity of compound translations as found in parallel corpora. This includes the discovery and extraction of noun compounds and the development of cross-lingually supported methods for analyzing noun compounds structurally (i.e., splitting of closed compounds and parsing of segmented/open compounds) and semantically (i.e., deriving the semantics of components and the relation in between). Possible collaboration partners include B1 (exploration of the multilingual noun compound database), B9 (integrating derivation and compound analysis), D8 (parsing models

for noun compounds) and SemRel (distributional models for the semantic interpretation of noun compounds).

Stefan Müller (D11)
Cross-lingual compound splitting for compound analysis

The D11 project adresses crosslingual compound splitting in WP2. Analyzing (ambiguous) compounds like "Wachstube" from a monolingual perspective often hits the wall, while crosslingual information can provide more insight for compound analysis: "guardroom" vs. "wax tube". Driven by this insight our current work deals with compound splitting as a precondition for this, including evaluation and results.
In my dissertation work in project D11 I adressed so far the topic of noun compound splitting with a multilingual, corpus-based approach. I developed a linguistically informed compound splitting tool which I attempt to improve at the moment. Currently, only English is employed to help splitting German compounds, but I plan to include further languages in future work. Furthermore, I plan to perform intrinsic (several GS formats are propagated) as well as extrinsic evaluation (within SMT-task). Our current work can be found at the first level of D11's tripartite, interdependent level approach, and is a prerequisite for the subsequent levels. In a later stage of my PhD research I will work on the two deeper levels of the model: determination of the constituents' meaning and uncovering implicit relations, focusing on methods that involve distributional semantics.
I see potential for methodological and contentual collaboration with projects also dealing with compositionality (e.g. B1, D12) or distributional information (e.g. B9, D10). Between B1/INF and D11 there's already an ongoing collaboration established (Noun Compound Database).

Britta Zeller (B9)
Induction, semantic validation and evaluation of a derivational morphology lexicon for German

We address the high language variability arising from German derivational morphology by inducing DErivBase, a lexicon of derivational families. It provides groups of derivationally related lemmas (i.e., surface-structure relations like “ask – asker – unasked - ...”), and is induced by combining a rule-based approach with corpus evidence. Since derivational relationship does not always imply semantic relationship, we refine DErivBase in a second step, distinguishing semantically transparent (“depart – departure”) and semantically opaque relations (“depart – department”). This refinement is conducted with supervised machine learning methods. We evaluate both lexicon versions on three extrinsic tasks, touching lexical semantics, psycholinguistics, and recognising textual entailment. We find that our derivational lexicon can help solving these tasks, and that the purely morphological and the semantically refined versions have different strengths and weaknesses.
Ideas for collaboration in the SFB: - D12: DErivBase as reference for particle verbs; indications for regularities in meaning shifts from our semantic refinement procedure - D11: Combination of two different morphological resources for German (the Compound Noun Database and DErivBase) to increase coverage for applicability - D2: Use DErivBase to restrict possible lexical content for sentence generation or improve text fluency .

Max Kisselew (B9)
Obtaining a better understanding of distributional models of German derivational morphology

My PhD project revolves around the modelling of different morphological processes using methods from distributional semantics. The current focus is on modelling the morphological process of derivation using two methods from compositional distributional semantics. We have already obtained promising results in a study on German derivational words and in another study on asymmetry in derivation. In an ongoing collaboration with the the more theoretically grounded project B4 we attempt to apply methods from distributional semantics to classify particle verbs and figure out underlying cognitive principles which might be helpful to discern the different verb classes. The development of human-readable representations of lexical and syntactic shifts in derivations is an essential topic in this collaboration, too.

Maximilian Köper (D12)
Modelling sense discrimination and regular meaning shifts of German particle verbs

When a base verb is combined with a particle, the resulting particleverbs often undergo meaning shifts. In this talk I will talk about this phenomenon, our (D12) hypothesis regarding meaning shifts and first experiments concerning a type-based classification of German particle verbs.

Silvia Springorum (D12, B4)
Shifting senses - from perception to German particle verbs

My dissertation aims to investigate possible senses and sense shifts of German particle verbs (PV) and PV neologisms, by not only taking lexical semantics into account, but also cognitive concepts, to shed light on less transparent PV compositions of, for example, metaphorical usages. An example is *aufbrummen,* meaning 'to burden s.o. with s.th.', which is a social interaction, although its base verb is from a completely different domain. I will present a psycho-linguistic experiment, which aims to identify possible directional concepts and their relations to PV compositions, by asking subjects to associate a set of visually presented arrow pictograms to systematically constructed PVs and their base verbs. The results show that there are agreements between particles and direction pictograms. The outcomes are a basis for the eye-tracking experiment designed by Evi Kiagia and will be theoretically analysed together with B4.

Evangelia Kiagia (D12)
Psycholinguistic processing of German particle verbs

In this talk I address the topic of psycholinguistic processing of German particle verbs (PVs). A series of experiments explores a) embodiment features in particle readings, b) the roles of particles and base verbs in the processing of the PVs, and c) whether information from adverbial adjuncts can serve as an indicator of metaphorical vs. literal contexts. In experiment 1, participants watch configurations of image schemas, while they listen to sentences: *"Der Stadtgärtner hat die Pflänzchen in kleinen Gruppen ausgegraben."* Eye movements towards the most relevant schema could confirm the relationship of PVs with embodiment features of

image schemas. In experiment 2 we measure reading times on divided versus complete particle verb formations, that can serve as predictors of PV meanings. Finally, in experiment 3 we assume that manner adverbs combined with PV arguments create a constraining context for the anticipation of literal vs. metaphorical PVs. The information gathered from the experiments serves on one hand, as empirical evidence on the processing of Pvs and on the other hand could be extended to support computational modelling studies.

Friday Talks

Despina Oikonomou (B6)
Middle (synthetic) vs. passive (analytic) voice: interpretation within or across cyclic domains

Crosslinguistically, we observe the following generalization:
(1) Synthetic Passives (e.g. Greek) can receive additional interpretations beyond the prototypical passive whereas analytic passives (e.g. English) cannot. We argue that this generalization is not accidental and that the different morphological properties of the Passives reflect different syntactic derivations. In the case of the English Passive, the Passive head creates a distinct interpretation domain thus blocking any other interpretation beyond the prototypical passive. On the contrary, synthetic Passive forms a single interpretation domain with the VP thus allowing a range of interpretations which depend on the properties of the VP predicate (cf. Alexiadou & Angonostopoulou 2004, Maranz 2007, 2013, Alexiadou & Doron 2012, Alexiadou 2013, Bobalijk & Wurmbrand 2013, Anagnostopoulou & Samioti 2013).

Réka Jurth (B6)
Resultative structures in Hungarian

The presentation focuses on Hungarian resultative constructions. I review the different types of resultative patterns with transitive, unaccusative and unergative verbs and discuss the types of resultative expressions (i.e. verbal particles, nominal expressions) that can occur in these constructions. Finally, I examine the behavior of experiencer verbs in resultative structures.

Doan Quyngoc (B8)
The Distribution of Vietnamese reflexive

Reflexivity is a topic that has drawn a lot of attention from linguistic researchers regarding generative grammar. In this study, we will provide a description of the distribution of the Vietnamese reflexive *mình*(self) based on the framework of Government and Binding Theory.

Nadja Schauffler (A7)
Prosody-inherent factors in L1 and L2 sentence production and perception

It is well known that information structure and prosody interrelate in many languages. For example, in German, information structure is typically marked by pitch accent placement and pitch accent type. However, the actual choice of pitch accent type as well as the distribution

of pitch accents in a phrase can vary beyond of what semantic or pragmatic factors can explain. In my dissertation I will look at how rhythmic factors interact with information structure in production and perception. In particular, I will investigate whether German and English speakers produce pragmatically induced accents (under contrastive focus conditions) when they violate rhythmic preferences (i.e. when two pitch accents would clash). I also plan to look at how this interaction is handled in non-native speaking and listening.
In collabroation with Kati Schweitzer (INF), we acoustically analysed German production data from a reading experiment applying the PaIntE-Model. This collaboration is planned to be continued to investigate other prominence-lending cues, like duration. Other aspects of collaboration could be:
• Testing the hypotheses on corpus-data
• Investigate how rhythm interrelates with other semantic or pragmatic aspects

John H.G. Scott (A7)
L2 acquisition of alternation and phonotactic constraints governing distribution of German [ç], [x], and [h] by L1 American English speakers

In German, [x] and [ç] are allophones (e.g., *Buch* [buːx] ~ *Bücher* ['byːˌçɐ]), and in German and English, [h] has a defective distribution ([huːt] "hat," but not *[tuːh] or *['aːˌtəhm]). Although position sensitivity is generally acknowledged as a factor, models of L2 phonological acquisition offer little to inform the formulation of testable hypotheses about more abstract aspects of phonological grammar (e.g., allophonic alternation, defective distribution). This project uses "ich-Laut" / "ach-Laut" alternation and the prosody-conditioned distribution of [h] in both German and English as test cases to shed light on adult L2 acquisition of higher-order phonotactic principles and narrow the gap between theory and empirical methods.

Fabian Bross
The iconic expression of scope orders of clausal categories in German Sign Language

This talk investigates the manual and non-manual encoding of high operators (e.g. speech-act operators, evaluative or modal operators) in German Sign Language. The central hypothesis is that high operators are encoded on a high place, i.e. suprasegmentally with facial expressions, and low operators further down, i.e. using the hands. In using the vertical axis as a scopally relevant dimension this encoding strategy mirrors iconically what cartographic approaches since Cinque (1999) assume: there seems to exist a cross-linguistic stable scope order of clausal categories.

Ina Rösiger (A6)
A pilot experiment on exploiting translations for literary studies on Kafka's Verwandlung

My PhD project is about the computational modelling of information structural aspects. The overall plan is to investigate the interaction between coreference resolution, bridging resolution and information status classification, mainly in German texts (DIRNDL and the new radio interviews), possibly also in English. As my project A6 also involves theoretical analyses of these data, I am interested in testing the theoretical claims in practical CL applications. In a recent paper, Arndt and I for example showed that including prosodic

information can help coreference resolution of spoken text. While the interaction between prosodic prominence and coreference has been investigated in experimental studies, it has never been tested in practice before.
I currently base most of my experiments on IMS HotCoref, our in-house coreference system provided by Anders (D8). We are planning on collaborating in future experiments wrt coreference and bridging resolution. I am also working closely together with Kerstin (INF) wrt DIRNDL and data questions, as well as Markus (INF) for ICARUS visualisation aspects. Together with Kati (INF) and Antje (A4), we have looked at the relation between prominence and coreference and we plan to collaborate further on automatic prosody labelling. I also have vague ideas on using distributional information for coreference and bridging resolution and would like to look into this with someone working in one of the distributional projects B9/D10/... .

Fritz Kliche (INF)
An online application for extracting text contents and metadata from text collections

A large amount of text data is available in digital form. I work on a web application which helps researchers in Digital Humanities projects to get access to the contents of these data. Users can import text data in several formats and data structures. The application offers methods for segmenting the text data into structural units and for extracting text contents and metadata. The resulting corpus can be cleaned from noise (corrupted and (near-)duplicate entries) and exported to different formats. The components of the web application are generic and can be used for processing different kinds of text collections. In my presentation, I will give an overview on the functions of the application and the integrated NLP methods. I will give some examples of text data which were processed by the application, which includes newspaper articles and tests with data from the INF project of the SFB. Finally, I will outline my future work, which includes the integration of named entity recognition into the application.

Abhijeet Gupta (D10)
Distributional vectors encode referential attributes

Distributional methods have proven to excel at concept based language-internal tasks like capturing fuzzy, graded aspects of meaning, like *Italy is more similar to Spain than to Germany*. While on one hand, distributional representations have been effectively used to represent general concept similarities, on the other hand they have not been considered fine-grained enough to represent specific referential attributes of a concept's instance, like *Italy has 60 million inhabitants*. We pursue the hypothesis that distributional vectors also implicitly encode referential attributes. In this study, we concentrate on Named Entities (where we have exactly one referent per concept) and show that these attributes can be retrieved to a reasonable degree of accuracy through a supervised regression model. We see this work as a small step towards bridging the concept-referent gap in distributional semantics.

Collaborations:
1. External Collaborators: Gemma Boleda, Marco Baroni
2. SFB Collaboration: Project D-11 (Compound Nouns)