Final Report AG 11

"Big Data: New opportunities and challenges in language acquisition research"

Language acquisition research on topics ranging from phonological processing to semantic knowledge has been built on meticulous examination of small data sets, such as single case studies. While we have learned a lot from such careful work, some limitations quickly became evident. A new horizon has opened as bigger, open data sets began to emerge. The workshop provided a platform for language acquisition researchers to assess the progress towards high quality, big, and open data sets, and to discuss solutions for current challenges. All information connected to the workshop, including slides of the presentations, can also be found on the workshop website at https://sites.google.com/site/chbergma/bigdata

Invited speaker Katherine Demuth opened the workshop by discussing one of the earliest examples of big, open datasets in child language acquisition research, namely corpora of natural speech, a prominent example being CHILDES (Child Language Data Exchange System), which contains openly available transcriptions of both children's' speech and their input and in many cases additionally audio files, videos, and annotations on different levels of granularity. On this platform, the speaker herself contributed three corpora, one in English, one in French, and one in Sesotho. These corpora together comprise over 800 hours of recordings centered around 17 children aged between one and four years. Katherine Demuth shared which opportunities arise from making such corpora available. Specifically, corpora are indispensable for understanding little studied languages or not yet systematically examined phenomena within a language, providing an exploratory basis for the subsequent design of more targeted experiments. Furthermore, many corpora track the development of children over a longer time span, providing a basis for our understanding of longitudinal language development and thus complementing snapshots obtained in laboratory experiments. Web-based platforms such as CHILDES have set an example for efficiently sharing open corpus data.

On the other side, there are also challenges connected with building corpora. On the one side, data collection can be difficult, both when working within the researcher's community (for example collecting data in Providence, USA) and when leaving the native environment to gain insight into previously under-studied languages (as going to South Africa to collect data on Sesotho). In both cases, data collection is limited by the environment (for example video might not be feasible if there is no readily available electricity) and the participants' level of consent and comfort. Transcription of collected data still has to be done by hand, which is time-consuming and thus very costly, as no reliable automatized solution is available at the moment.

To conclude, Katherine Demuth shared recommendations, such as aiming for recording videos whenever possible, sharing work because others will have questions and ideas that are independent of the initial purpose and still can be answered with the same corpus, to collect a monolingual corpus before going on to answer the same question in a bilingual dataset, and that cross-linguistic research questions can best be answered using ideally parallely constructed corpora.

The second presentation (Lavalley, Berkling, & Stueker) introduced a web-based platform to make a corpus of texts written by children in grades 1-8 accessible and to provide analyses of spelling error analyses. The goal of these corpora is to identify points of improvement for German spelling teaching in the school-environment and to identify commonly made errors, a research question impossible to answer without using corpora. One of the methodological conclusions - which is relevant across subfields in Linguistics - is the relevance of the prompts used to elicit any form of data. This topic is currently understudied according to the speaker. The data gathered from an interactive web interface was automatically annotated using speech synthesis tools to classify spelling. The database can be accessed by researchers, and data can be added by users. To motivate contributions the automatic spelling error analysis is provided after submission.

The third presentation (Hills & Yoshida) described how the authors used large, already available datasets to predict the order of word learning in monolingual and bilingual first language learners based on the associative structure of words. This approach relates children's productive vocabularies, adult norms of free associations, and the statistical structure around words derived from corpora of child-directed speech. One central insight from this work (utilizing data from around 400 children) is that monolingual children tend to learn words that are more contextually diverse in both adult and child-directed language, with bilingual children showing an amplification of this tendency.

The following two projects described longitudinal infant studies in which different types of data were collected at multiple time points.

McGillion, Herbert, Pine, and Matthews presented an intervention study to test the causal role of caregiver contingent talk on a large sample of 150 eleven-month-old infants' language learning. Contingency, the time-locked and topically related reaction of a caregiver to a child's utterance, might account for some of the difference found between levels of socioeconomic status. However, a causal link has not yet been established and to this end, a longitudinal intervention study which collects a lot of data from each infant was devised. The contingency before the intervention as well as the outcomes were assessed using audio and video recordings and language questionnaires before the intervention and at various time-points after the intervention. Furthermore, infants were tested in laboratory tasks on their cognitive capabilities. Preliminary results suggest that a contingent talk training lead to a higher amount of caregiver contingent talk and to higher infant vocabulary scores later on, independent of caregiver socioeconomic status.

Dijkstra, Benders, and Fikkert discussed data management in their longitudinal Place Perception Production (PPP) project, which investigates early speech perception, early speech production, and the developmental relationship between perception and production for the place of articulation of consonants. Fifty-six Dutch infants at the ages of eight, twelve and sixteen months participated in perception experiments and home recording sessions, and their parents filled out a battery of questionnaires.

Finally, Tsuji, Bergmann, and Cristia introduced community-augmented meta-analyses (CAMAs), a combination of classical meta-analysis and an freely available and continuously updated online repository. Meta-analysis of extant data can not only provide an overall picture of the presence, size, and variance of an effect and its moderators, but also help in planning new studies, namely by enabling power analyses, stimulus selection, and more. By making such meta-analyses available and updatable by users, resources can be shared, and databases can stay up to date.

Despite the diversity in topics and approaches in this workshop, the discussion of benefits and challenges of relatively large datasets in language acquisition research converged. All presenters agreed that the size and complexity of the datasets they used or collected was necessary for answering their research questions - by removing the noise from the signal, by providing the necessary power, or by enabling a multifactorial view on a research topic. Challenges perceived were also multifold. Large datasets require a good data management and storage strategy. We discussed the advantages and disadvantages of automatic versus manual solutions to data annotation and management, issues with anonymization, data protection and privacy when sharing data as well as appropriate statistical models and data subsample selection to analyze complex relationships.

We concluded that increased sharing of information across laboratories who plan to embark on a project involving big data will be beneficial for addressing these challenges.