Workshop "Perspectives on low-resource language varieties", February 9th 2018
Language Science and Technology, SFB 1102

Birgit Hellwig (University of Cologne)
Child language documentation: Perspectives from language documentation, language acquisition and language socialization.

In this talk, I discuss possibilities and challenges of child language documentation, aiming to integrate three different perspectives: language documentation, language acquisition (psycholinguistics) and language socialization (anthropology). Despite a general recognition that our databases, and hence our theories, are severely skewed towards WEIRD languages and societies (e.g. Anand et al. 2011; Henrich et al. 2010), there are only very few initiatives that attempt to broaden our databases and to include low-resource languages. This talk discusses possible ways of addressing this bias, focusing on a case study: the documentation of child language among the Qaqet of Papua New Guinea.

Alena Witzlack (University of Kiel)
Differential argument marking: from linguistic typology to corpus studies and back.

In many languages of the world an argument of the same verb, e.g. its object, can carry various case marking. Frequent triggers of such variation are e.g. animacy or definiteness of the respective argument. This phenomenon is known as *differential object marking* or *DOM* (Bossong 1985). Many languages show similar patterns of variation which pertain to other arguments, affects agreement and are conditioned by factors other than animacy or definiteness. Such phenomena are collectively referred to as *differential argument marking*. In this paper, I will first introduce a set of variables which enable typological studies on differential argument. This set of variables should ideally allow for straightforward comparison across languages and capture the fine differences between them. Several cases studies are presented to illustrate how such a system of variables works.

Most of the earlier research on differential argument marking assumed that rule-based descriptions listing one or two factors are sufficient to predict the distribution of the respective markers correctly. Which exactly factors these are is often less clear in individual cases, cf. for instance, the many studies dedicated to the factors conditioning the distribution of *a*-marked vs. unmarked direct objects in such a well-studied language as Spanish (e.g. Hopper & Thompson 1980, Brugè & Brugger 1996, Aissen 2003, von Heusinger & Kaiser 2007). A number of recent studies argue that probabilistic models of differential argument, which can quantify the impact of many factors on the form, allow for a much better prediction of the correct form than rule-based description (e.g. Schikowski 2013). Such probabilistic models are fitted on the bases of large annotated corpora and require more analytical work than traditional rule-based accounts. The second part of the presentation addresses some of the challenges related to applying such models for cross-linguistic research and outlines a pilot project aimed to investigate the distribution of differential object agreement in a number of Bantu languages.

References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic theory* 21(3). 435–483.
Bossong, G. 1985. *Differentielle Objektmarkierung in den neuiranischen Sprachen.* Tübingen: Narr.
Brugè, L. and Brugger, G. 1996. On the accusative a in Spanish. *Probus* 8: 1–51.
Hopper, P. & Thompson, S. 1980. Transitivity in Grammar and Discourse. *Language* 56: 251–299.

Schikowski, R. 2013. *Object-conditioned differential marking in Chintang and Nepali.* Zurich: University of Zurich PhD thesis.
von Heusinger, K. & G. A. Kaiser. 2007. Differential object marking and the lexical semantics of verbs in Spanish. In G. A. Kaiser & M. Leonetti (eds.), *Proceedings of the workshop "Definiteness, specificity and animacy in Ibero-Romance languages"*, 83–109. Universität Konstanz: Fachbereich Sprachwissenschaft.

Alexander Koplenig (Institute for the German Language (IDS), Mannheim)

Languages with more speakers tend to be harder to learn

Are all languages equally hard to learn? This contribution presents the results of a large scale computational experiment, where an "ideal learner" (Chater & Vitányi 2007) learns more than 1,100 different languages.

In the first part, I will discuss the idea that (much of) cognition can be understood as compressing data, i.e. finding a description of the (sensory) input that is as short as possible, both (i) in terms of the elimination of redundancy and (ii) in order to learn to predict subsequent data based on the preceding input (Chater & Vitányi 2003). Interestingly, the entropy rate as defined by Shannon (1948; 1951) quantifies both (i) and (ii).

In the second part, I will pick up the idea of Jamison & Jamison (1968), who suggest to examine language learning performance by measuring the decrease of the entropy rate as a language is learned. To this end, an efficient Markov-based compression algorithm is exposed to increasing amounts of written linguistic data. Language learning performance is measured by statistically estimating the speed of convergence of the compression rate to the entropy rate (Takahira et al. 2016). This procedure is applied to parallel translations of the Bible (Mayer & Cysouw 2014). It is shown that (i) languages with more speakers tend to be more complex and – statistically independent of (i) – that (ii) languages with more speakers tend to be harder to learn. I demonstrate that observation (ii) holds for different language families and for different geographical areas and is still valid after controlling for a set of possible confounding variables.

In the last part of my talk, I try to isolate and identify potential mechanisms that can help to understand this relationship.

Chater, Nick & Paul Vitányi. 2003. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences* 7(1). 19–22. doi:10.1016/S1364-6613(02)00005-0.
Chater, Nick & Paul Vitányi. 2007. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51(3). 135–163. doi:10.1016/j.jmp.2006.10.002.
Jamison, Dean & Kay Jamison. 1968. A note on the entropy of partially-known languages. *Information and control* 12(2). 164–167.
Mayer, Thomas & Michael Cysouw. 2014. Creating a Massively Parallel Bible Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3). 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
Shannon, Claude E. 1951. Prediction and Entropy of Printed English. *Bell System Technical Journal* 30(1). 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x.
Takahira, Ryosuke, Kumiko Tanaka-Ishii & Łukasz Dębowski. 2016. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* 18(10). 364. doi:10.3390/e18100364.

Yunfan Lai and *Johann-Mattis List* (MPI-SHH, Jena)

Investigating Verb Derivation Patterns in Sino-Tibetan Languages within a Computer-Assisted Framework

The ERC-funded research project *CALC* (Computer-Assisted Language Comparison, http://calc.digling.org) establishes a computer-assisted framework for historical linguistics. We pursue an interdisciplinary approach that adapts methods from computer science and bioinformatics for the use in historical linguistics. While purely computational approaches are common today, the project focuses on the communication between classical and computational linguists, developing interfaces that allow historical linguists to produce their data in machine readable formats while at the same time presenting the results of computational analyses in a transparent and human-readable way.

In particular, we are interested in historical patterns of verb derivation in Sino-Tibetan languages. Here we intend to carry out a pilot study on the diachronic aspects of verb alternation in Rgyalrongic languages (especially Khroskyabs and Stau) which will eventually serve as a basis for an etymological dictionary of the Sino-Tibetan subgroup. In contrast to previous approaches which are largely anecdotal, listing examples and supposed derivation pathways with semantic explanations, we pursue an onomasiological approach which starts from a pre-compiled list of fundamental semantic alternations of a list of basic verb meanings with semantically annotated source and target meanings (e.g., eat → feed, sit → seat, etc.) which we then annotate etymologically, using annotation frameworks and tools currently developed in our project.

In the talk, we will briefly present the general framework of Computer-Assisted Language Comparison along with its basic tools for data annotation and analysis and then explain how we intend to tackle the problem of verb alternations in Rgyalrong in specific. Since this research is in its initial stage, we hope to instigate a more vivid discussion on the usefulness of our annotation ideas compared with the problems which scholars face in other fields of historical linguistics and language typology.

Dietrich Klakow (Saarland University)

Low-resourced?

When talking about low resource languages many people think about endangered languages. However, we have a much bigger "blind spot": for about 300 languages there are more than 1 million speakers but still no speech or language technology exists for most those languages. In this talk I will introduce methods for

* vocabulary construction for speech recognition systems
* calculating embeddings and
* named entity recognition

targeting those languages.