Human Language Processing / Jaeger Lab

Contact | Calendar | People | Participate in an Experiment | More about HLP Lab research | Job & Research Opportunities | Lab Blog | Resources

Empirically examining parsimony and redundancy in usage-based models

Edited Volume/Special issue (in prep.)

Editors:

Neal Snider, Nuance Communications, Inc.
Daniel Wiechmann, RWTH-Universität Aachen
Elma Kerz, RWTH-Universität Aachen
T. Florian Jaeger, University of Rochester

Contact: wiechmann@anglistik.rwth-aachen.de and kerz@anglistik.rwth-aachen.de

This year's LSA Annual Conference featured a workshop on Empirically examining parsimony and redundancy in usage-based models organized by Neal Snider (Nuance Communications, Inc.), Daniel Wiechmann (RWTH Aachen University), Elma Kerz (RWTH Aachen University) and T. Florian Jaeger (University of Rochester. The workshop brought together linguists, psycholinguists, and computational linguists and was geared to discuss which methodologies can best shed light on questions pertaining to the representational nature of constructions and the mechanisms involved in their on-line processing. The workshop abstract is given below.

Based on this workshop, we are currently preparing an edited volume, which represents the results of the workshop but also comprises additional work in this domain. Thus the objective of the volume is to present the state-of-the-art of

research into parsimony and redundancy in usage-based models drawing from a variety of methodologies and interdisciplinary approaches. To this end, we welcome papers that empirically examine these issues at any level of linguistic description.

Workshop abstract:

Recent years have seen a growing interest in usage-based (UB) theories of language, which assume that language use plays a causal role in the development of linguistic systems over historical time. A central assumption of the UB-framework is the idea that shapes of grammars are closely connected to principles of human cognitive processing (Bybee 2006, Givon 1991, Hawkins 2004). UB-accounts strongly gravitate towards sign- or construction-based theories of language, viz. theories that are committed to the belief that linguistic knowledge is best conceived of as an assembly of symbolic structures (e.g. Goldberg 2006, Langacker 2008, Sag et al. 2003). These constructionist accounts share (1) the postulation of a single representational format of all linguistic knowledge and (2) claim a strong commitment to psychological plausibility of mechanisms for the learning, storage, and retrieval of linguistic units. They do, however, exhibit a considerable degree of variation with respect to their architectural and mechanistic details (cf. Croft & Cruse 2004).

A key issue is the balancing of storage parsimony and processing parsimony: Maximizing storage parsimony is taken to imply greater computational demand and vice versa. The space of logical possibilities ranges from a complete inheritance model (minimal storage redundancy) to a full-entry model (maximal storage redundancy). Currently, the empirical validation of the theoretical situation is not yet conclusive: the representations involved in language processing involve extremely fine-grained lexical-structural co-occurrences, for example frequent four-word phrases are processed faster than infrequent ones (Bannard and Matthews 2008, Arnon and Snider 2010). On the other hand, syntactic exemplar models (Bod 2006) have been argued to overfit and undergeneralize compared to models that do not store all structures in the training data (cf. Post and Gildea 2009, although they found that Tree Substitution Grammar representations induced in a Bayesian framework still split the parsimony continuum towards greater redundancy). Also, experimental work has argued that models of categorization that directly map phonetic dimensions to phonological categories (and therefore more directly reflect the statistics of the training data) do not predict human behavior as well as models that assume independent, intermediate representations (Toscano and McMurray 2010). Additionally, recent work has provided evidence that early evidence for full-entry models from item-based learning in acquisition (e.g. Pine & Lieven 1997) is confounded, reopening this line of research as well (Yang, unpublished manuscript).

References and representative work

- Arnon, I. and N. Snider. 2010. More than words: Frequency effects for multi-word phrases. Journal of Memory and Language. 62, 67-82.
- Bannard, C. and Matthews, D. 2008. Stored word sequences in language learning: The effect of familiarity on childrens repetition of four-word combinations. Psychological Science, 19, 241248.
- Bod, R. 2006. Exemplar-based syntax: How to get productivity from examples. The Linguistic Review 23.291320.
- Bybee, Joan. 2006. From usage to grammar: the mind's response to repetition. Language 82(4). 711-733.
- Croft, W and D. A. Cruse. 2004. Cognitive linguistics. Cambridge: Cambridge University Press.
- Feldman, J., E. Dodge and J. Bryant. 2009. A Neural Theory of Language and Embodied Construction Grammar. In The Oxford Handbook of Linguistic Analysis. Edited by B. Heine and H. Narrog.
- Goldberg, A. 2006. Constructions at work: The Nature of Generalisation in Language. Oxford: OUP
- Hawkins, J.A. 2004. Efficiency and Complexity in Grammars. Oxford University Press, Oxford

- Konopka, A. E., &Bock, J. K. 2009. Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. Cognitive Psychology, 58, 68101.
- Langacker, R. 2008. Cognitive Grammar: A Basic Introduction. New York: Oxford University Press
- Post and Gildea (2009) Post, M. and D. Gildea. 2009. Language modeling with tree substitution grammars. Proceedings of NIPS workshop on Grammar Induction, Representation of Language, and Language Learning. Whistler, British Columbia.
- Sag, I. A., T. Wasow, and E. Bender. 2003. Syntactic Theory: A Formal Introduction (Second Edition). Stanford: CSLI Publications.
- Tomasello, M. 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press.
- Toscano, J. C., & McMurray, B. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science, 34, 434-464.

Accepted papers:

- Antal van den Bosch (Tilburg University/University of Antwerp) & Walter Daelemans (University of Antwerp) - Implicit schemata and categories in memory-based language processing
- Daniel Gildea (University of Rochester) and Matt Post (University of Rochester) - Sampled Representations for Tree Substitution Grammar
- Timothy John O'Donnell (Harvard University) - Productivity and Reuse in Language
- Harald Baayen (University of Alberta); Marco Marelli (University of Alberta); Peter Hendrix (University of Alberta) and Petar Milin (University of Alberta) - Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning
- Ed Holsinger (University of Southern California) - Representing Idioms: Syntactic and Contextual Effects on Idiom Processing