*Proceedings of AESPEN 2020*, page 8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# Keynote Abstract: Current Open Questions for Operational Event Data

**Philip A. Schrodt**
Parus Analytics, LLC
Charlottesville, Virginia, USA
schrodt735@gmail.com

*Abstract content*

In this brief keynote, I will address what I see as five major issues in terms of development for operational event data sets (that is, event data intended for real time monitoring and forecasting, rather than purely for academic research).

First, there are no currently active real time systems with fully open and transparent pipelines: instead, one or more components are proprietary. Ideally we need several of these, using different approaches (and in particular, comparisons between classical dictionary- and rule-based coders versus newer coders based on machine-learning approaches).

Second, the CAMEO event ontology needs to be replaced by a more general system that includes, for example, political codes for electoral competition, legislative debate, and parliamentary coalition formation, as well as a robust set of codes for non-political events such as natural disasters, disease, and economic dislocations.

Third, the issue of duplicate stories needs to be addressed – for example, the ICEWS system can generate as many as 150 coded events from a single occurrence on the ground – either to reduce these sets of related stories to a single set of events, or at least to label clusters of related stories as is already done in a number of systems (for example European Media Monitor).

Fourth, a systematic analysis needs to be done as to the additional information provided by hundreds of highly local sources (which have varying degrees of varacity and independence from states and local elites) as opposed to a relatively small number of international sources: obviously this will vary depending on the specific question being asked but has yet to be addressed at all.

Finally, and this will overlap with academic work, a number of open benchmarks need to be constructed for the calibration of both coding systems and resulting models: these could be historical but need to include an easily licensed (or open) very large set of texts covering a substantial period of time, probably along the lines of the Linguistics Data Consortium Gigaword sets; if licensed, these need to be accessible to individual researchers and NGOs, not just academic institutions.