**Master class:** [Carla Parra Escartín](http://www.uib.no/persons/Carla.Parra),

**Title: Nominal compounds in German and Spanish**

UiB, Bergen. September 10th, 2014.

**Participants**: Dr. [Preslav Nakov](http://qcri.qa/our-people/bio?pid=35&par=acc&name=PreslavNakov), Qatar Computing Research Institute and PhD candidate [Carla Parra Escartín](http://www.uib.no/persons/Carla.Parra), University of Bergen.

In addition to the master class, Preslav Nakov will give a guest lecture.

Organized by the Norwegian Graduate Researcher School in Linguistics and Philology.

Please sign up by email to [Bamba Dione](mailto:dione.bamba@lle.uib.no) by Sept. 8 so that we may order coffee and refreshments.

The Master class and the guest lecture will be held in English.

**Program**

**Wednesday Sept. 10**

**Place**: Sydneshaugen skole, room 304B

13-14:30 Master class  
14:30-15 Break  
15-16:30 Master class  
  
**Thursday Sept. 11**

**Place**: HF-building, room 216

10:15-12 Guest lecture by Preslav Nakov

**The Web as an Implicit Training Set: Application to Noun Compounds Syntax and Semantics**

The 60-year-old dream of computational linguistics is to make computers capable of communicating with humans in natural language. This has proven hard, and thus research has focused on sub-problems. Even so, the field was stuck with manual rules until the early 90s, when computers became powerful enough to enable the rise of statistical approaches. Eventually, this shifted the main research attention to machine learning from text corpora, thus triggering a revolution in the field.  
  
Today, the Web is the biggest available corpus, providing access to quadrillions of words; and, in corpus-based natural language processing, size does matter. Unfortunately, while there has been substantial research on the Web as a corpus, it has typically been restricted to using page hit counts as an estimate for n-gram word frequencies; this has led some researchers to conclude that the Web should be only used as a baseline.  
  
In this talk, we will reveal some of the hidden potential of the Web that lies beyond the n-gram, with focus on the syntax and semantics of English noun compounds. First, we will present a highly accurate lightly supervised approach based on surface markers and linguistically-motivated paraphrases that yields state-of-the-art results for noun compound bracketing: e.g., "" is right-bracketed. Second, we will present a simple unsupervised method for mining implicit predicates that can characterize the semantic relations holding between the nouns in noun compounds, e.g., "malaria mosquito" is a "mosquito that carries/spreads/causes/transmits/brings/infects with/… malaria''. Finally, we will show how these ideas can be used to improve statistical machine translation.