Text processing of social media data in NLP has largely centred around Twitter, with other social media types getting relatively little attention. In this talk, I will first present an empirical comparison of text sourced from a range of social media sites (microblogs, comments, user forums, blogs and Wikipedia), focusing on the relative “noisiness” and diversity of the linguistic content. I will then discuss the application of discourse parsing to user forum threads, and incorporation of the results to improve search quality.

Tim Baldwin is a Professor in the Department of Computing and Information Systems, The University of Melbourne, an Australian Research Council Future Fellow, and a contributed research staff member of NICTA Victoria. He has previously held visiting positions at the University of Washington, University of Tokyo, Saarland University, NTT Communication Science Laboratories and National Institute of Informatics. His research interests include text mining of social media, computational lexical semantics, information extraction and web mining, with a particular interest in the interface between computational and theoretical linguistics. Current projects include web user forum mining, text mining of Twitter, and machine translation evaluation. He is currently Secretary of the Australasian Language Technology Association and a member of the Executive Committee of the Asian Federation of Natural Language Processing, and was recently PC Chair of EMNLP 2013 and *SEM 2013.

Tim completed a BSc(CS/Maths) and BA(Linguistics/Japanese) at The University of Melbourne in 1995, and an MEng(CS) and PhD(CS) at the Tokyo Institute of Technology in 1998 and 2001, respectively. Prior to joining The University of Melbourne in 2004, he was a Senior Research Engineer at the Center for the Study of Language and Information, Stanford University (2001–2004).

http://www.cs.cmu.edu/~nlp-lunch/