

Machine-aided manual retrieval and annotation of constructions: Desiderata for investigating grammatical variation in corpora

**Elena Seoane-Posse (University of Santiago de Compostela) &
Nicholas Smith (University of Salford)**

In the last few decades there have been significant advances in many areas of Natural Language Processing (NLP), e.g. in automated approaches to part-of-speech (POS) tagging, morphological analysis, speech recognition and text summarization. English, moreover, is the language where most research effort has been expended, and successful, sometimes commercially viable, systems developed. Despite such progress, however, the retrieval and coding of many, perhaps most, grammatical structures is not something that can currently be achieved to the level of accuracy and specificity that most linguists require in order to make robust descriptions. As a result, we still have to do a considerable amount of manual spadework ourselves just to be able to retrieve all – and only – valid instances of a given structure. Once those instances have been retrieved, there remains the still more difficult task of coding all the features we deem potentially relevant to describing and interpreting the use of that structure.

The aim of our paper is to address these two issues, the optimal retrieval and categorization of grammatical variants in a corpus. As a case study we focus on the analysis of long passives (i.e. passives with an overt agent phrase) and transitive actives in two contrasting registers of the British National Corpus. We will show that, even though these constructions are not pre-annotated in the BNC, they can be retrieved with acceptable levels of recall and precision with the aid of (a) automatic POS tagging provided in the corpus; (b) sophisticated query software such as *BNCweb CQP edition* (Hoffman et al. forthcoming); (c) use of other hand-coded corpora as training data to refine the queries.

Our discussion will then focus on the classification of long passives and transitive actives, discussing the possibilities and constraints of a computer-assisted manual approach, and how it can best be applied to identify factors underlying the choice between the respective variants.

Reference:

Hoffmann, S., S. Evert, N. Smith, D. Lee and Y. Berglund Prytz. (Forthcoming). *Corpus Linguistics with BNCweb—a Practical Guide*. Frankfurt: Peter Lang.