Quantitative Computational Syntax

Paola Merlo
University of Geneva

In the computational study of intelligent behaviour, the domain of language is distinguished by a complex and sophisticated domain theory and by large amounts of observational data in many languages. The main current scientific challenge for computational linguistics is the creation of theories and methods that fruitfully combine large-scale, corpus-based approaches with the linguistic depth of more theoretical methods. In these lectures, I will introduce some recent and current work from our group, where large-scale, data-intensive computational modelling techniques are used to address theory-driven linguistic questions.

Lecture 1: Motivations, data and methods.

This lecture will be organised in three parts:

1. Computational modelling: the third way to science
   In which we explain what a model is; we discuss the added value of modelling to the language sciences, comparing it to experimentation; and conclude with a brief summary of the debate on the place of statistical models in linguistics.
2. Data
   In which we discuss data from different points of view: expressiveness of different kinds of data; different data collection methods. We then summarise two debates on data: the debate on lack of quantitative data in linguistics and the dichotomy between curated data and uncurated data, data "in the wild". We then spend some time on linguistic distributions in corpora, powerlaw distributions, such as Zipf's law and report one example from child language acquisition to show that it is important to keep the actual distribution of the frequencies in mind in theorizing about language.
3. Methods
   In which we provide elements of methodology and models that are needed to understand the following lectures: for example, the machine learning paradigm, and notions such as probabilistic modelling and latent variables. We briefly discuss model performance and model comparison.

Lecture 2: The Grammatical Basis of Word Order Frequencies

In this lecture, we investigate the factors that govern one of the most apparent sources of diversity across languages: the order of words. First we report on work that investigates whether typological frequencies are systematically correlated to abstract syntactic principles at work in structure building and movement. Then, we investigate higher level structural principles of efficiency and complexity: the availability of several large-scale treebanks allows us to ask this question in a novel way. In a large-scale, computational study on Romance languages, we confirm a trend towards minimisation of the distance between words across languages even for short spans. These results raise issues on the role of efficiency and complexity in language use.

Lecture 3: Corpus-driven Accounts of the Causative Alternation

In this lecture we discuss the causative alternation. We first revisit previous work that shows that argument structure properties can be detected in a corpus and reliably used to identify lexical semantic properties that distinguish the causative alternation from other lexical semantic classes. Then, we show how, much like the comparative method in linguistics, cross-lingual corpus investigations take advantage of any corresponding annotation or linguistic knowledge across languages. We show that corpus data and typological data involving the causative alternation exhibit interesting correlations explained by the notion of inner causation of an event. This line of work leverages both similarities and differences across languages of the world.

Bio

Paola Merlo is associate professor in the Linguistics department of the University of Geneva. She is the head of the interdisciplinary research group Computational Learning and Computational Linguistics (CLCL). The group is concerned with interdisciplinary research combining linguistic modelling with machine learning techniques. Prof. Merlo is the current editor of *Computational Linguistics*, published by MIT Press and is a member of the executive committee of the ACL. Prof. Merlo studied theoretical linguistics at the University of Venice and holds a doctorate in Computational Linguistics from the University of Maryland, USA. She has been associate research fellow at the University of Pennsylvania, and visiting scholar at Rutgers, Edinburgh, and Stanford.