

## **Methods in Formal Linguistics**

Thursday March 18th, 2010

9 a.m. – 12 noon.

Teacher: Matthias Buch-Kromann

Department of International Language Studies and Computational Linguistics,  
Copenhagen Business School

This lecture focuses on Chomsky's thesis that natural languages are describable as *formal* systems (formal syntax), and Montague's thesis that natural languages are describable as *interpreted* formal systems (model theoretic semantics).

Our point of departure is the question of what human language is. We briefly discuss Chomsky's mentalist approach to the study of human language and his reason for postulating an innate faculty of language (FL) separate from general learning devices in the brain. FL is supposed to help infants overcome the problem posed by the varying quality and insufficiency of the linguistic input they encounter when developing their first language (the so-called "argument from the poverty of the stimulus"). Recently, Chomsky and others have proposed that there may not be very much structure in the FL, and that the decisive component of the FL is simply the ability to produce and understand infinitely many new utterances.

These preliminaries raises questions of what a linguistic theory is about and what the structure of a linguistic theory is, and how it constructs models of linguistic phenomena which are empirically testable. In this connection, of course, we have to take a look at what constitutes appropriate data for our theory, and if and how this data is available to the practising linguist.

We present some formal methods in linguistics, and look at what a formal grammar is and how such grammars may help us model linguistic data. Among the properties of formal grammars we note their recursive properties, and we demonstrate their usefulness in describing some non-trivial syntactic phenomena in English and Danish, such as longdistance dependencies and genitive constructions, which, incidentally, constitute examples of potentially infinite strings, cf. what we said about FL above.

Syntax and semantics will take centre stage in the lecture, and we will show how syntactic structure may serve as a skeleton on which to build semantic representations of sentences compositionally. It is emphasized that semantic representations are not just arbitrary paraphrases of natural language expressions, but are symbolic representations enabling strict, unambiguous and transparent interpretations of the natural language expressions whose meanings they represent.

Finally, we look at some practical perspectives in applying formal methods to the description of natural language, in particular in computer systems such as natural language interfaces to databases.

## Methods in Computational Linguistics

Thursday March 18th, 2010

1 p.m. - 4 p.m.

Teacher: Matthias Buch-Kromann

Department of International Language Studies and Computational Linguistics,  
Copenhagen Business School

In 1955, the American linguist C.F. Hocket thought that grammaticality represents a linguist's schematized notions of probability, where low probabilities correspond to "ungrammatical", and high probabilities correspond to "grammatical." In his 1957 book "Syntactic Structures", Chomsky explicitly rejected this idea and used the example "Colorless green ideas sleep furiously" to argue that grammaticality was completely unrelated to probability. Chomsky's research programme was so successful that probabilities did not reappear as a "hot" topic in linguistics before the 1990s, where the dynamics of the field suddenly changed because of the advent of cheap computers with unprecedented capabilities for data processing, and Fred Jelinek from the IBM Speech Laboratory was widely quoted for saying that "Every time I fire a linguist, system performance goes up" - a great sound-bite, although Jelinek actually never said it.

Today, the probabilistic approach has become the dominant paradigm in computational linguistics, and the rise of probabilistic methods has led to significant and continuing improvements in the best systems for machine translation, speech recognition, information extraction, and syntactic analysis. However, there is a growing sense in the field that probabilities on their own are not enough, and that we may be about to witness an interesting convergence between linguistics, computational linguistics, and psycholinguistics and brain science in the coming 5-10 years.

In this lecture, I will focus on how the probabilistic methodology has led to a division of labour between linguists and computational linguists: the linguists create linguistically annotated data collections consisting of large amounts of text annotated with various levels of structure (syntax, morphology, phonology, rhetorical structure, anaphora, semantics, and the structure of parallel translations); and the computational linguists create computational models that can exploit the information in the annotated data on the basis of the linguistic intuitions that are encoded in the models. The result is a set of hugely complex probabilistic language descriptions that are far better than anything we could have created by hand.