Overview of DAAD/PPP project: “Natural Language Ontology for Reference to Facts and Events”
Cooperation Partner: University “Pompeu Fabra”, Barcelona

Natural languages provide various means of referring to abstract objects such as facts, events, propositions, actions, states and processes. Our goal is to develop a robust ontology for these types of objects which supports semantic analysis in a cross-linguistic setting and provides the foundation for semantic-aware computational applications. While these kinds of abstract entities have been observed to have different properties which have often been the subject of philosophical investigation (Kim, Davidson), little progress has been made in developing criteria for distinguishing when a linguistic expression refers to one or the other of these objects, nor is it even entirely clear what the complete inventory of such abstract objects is. Nonetheless, such criteria and a natural language ontology to support linguistic analysis are essential to understanding the similarities and differences in the ways different languages regulate reference to abstract objects; such understanding, in turn, is essential not only for the insights it offers into the nature of human language but also for very practical natural language processing applications such as automatic translation or the resolution of anaphoric expressions referring to such entities.

The goal of this project is to bring together researchers with backgrounds in three related areas - linguistic (particularly semantic) theory, computational semantics/discourse, and corpus linguistics – to investigate a focused set of questions concerning the natural language ontology of abstract objects as it relates to nominals and anaphora. We consider this combination of methodologies and perspectives crucial to making progress in this area, and expect the project to be a test-bed for such collaboration.

The fact/eventuality domain is of interest because it has played an important role in a number of as yet only loosely related domains. There has been considerable work in this domain in the philosophical tradition (e.g. Bennett, Asher), but the application of the insights gained here to the analysis of the linguistic manifestations of these objects, particularly in a cross-linguistic perspective, has been very limited (e.g. Chierchia, Zucchi). There is, however, a substantial body of research on the relationship between lexical aspect (Aktionsart) and verbal argument structure, and between lexical aspect and tense and sentence aspect. The typology of events used here is clearly not well enough articulated to facilitate an adequate treatment of many other linguistic phenomena, though, such as the restrictions on the distribution of nominal predicative expressions with the range of copular verbs and the range of objects denoted by verb-phrase anaphora, two empirical areas we will investigate in detail.

We expect that the study of nominal, nominalization processes, and anaphora will entail further refinement of this well-understood typology of events. Additionally, the tracking of referents in discourse has become an important area for application development within computational linguistics. Computational discourse is now a well-developed field. This existing work offers a solid base from which we can explore an understudied empirical domain. Finally, the ontology necessary for analysing nominal reference to abstract objects will also be essential for the analysis of other grammatical categories, most notably for the treatment of adjectives and verbs, making the work we do in this initial project an excellent stepping stone to future projects involving these other categories.

Specific Objectives

Our objective for the period 2006-2008 is the development of an ontology of abstract objects that facilitates the semantic analysis of two related domains: nominalization and nominal argument structure on the one hand and verb phrase anaphora (e.g. 'do so', 'do it', Ger. 'dabei', 'das') and neuter pronominal anaphora on the other (e.g. 'it', 'this', Cat. 'ho', Ger. 'es'/das'). Our work will be based on data from Catalan, Spanish, English French and German, and the theoretical linguistic analysis will be supported by and evaluated through corpus-linguistic analysis. With respect to nominalization, two phenomena will be the subject of detailed investigation: nominalization type and nominal argument type. We will investigate, for
example, why it is natural to say the solution of the problem took over two hours and the solution to the problem is on page 24, but not the other way around (likewise in German and Spanish). In addition, we plan to investigate the various ways in which different nominalization processes (conversion; derivational suffixes such as -tion/-ción, -ité/-ity/-dad, Eng. -ness; -ing nominals and German -ung and -en nominals) correlate with different semantics for the resulting derived nominals (Badia & Saurí 2001, Colominas 2001).

With respect to anaphora, we will investigate the different types of abstract objects referred to by various anaphoric and deictic forms. For example the account for the contrasts between it and this and do it in (1) will be of central interest for the development of the ontology:

(1) Sandy devoured the cake. It took only minutes./That disturbed us./She did it quickly.

(see Cornish 1999 for a recent general study; see also Fontana 2000 on the specific choice between it and this, and e.g. Bosch 1983 on the distinction between discourse deixis and anaphora). Ongoing work by Miller on the complements to perception verbs also sheds light on this question, as the ontological type of these verbs’ complements (entity, event, fact, state) is crucial for their syntax. In addition to our theoretical goals, we have two computational goals, the first being the generation of a distributional analysis of the various forms of nominalization and anaphora in the languages under study and the second being the generation of a preliminary annotated corpus. The annotation here envisioned would involve classifying relevant utterances in terms of the type of abstract entity referred to. By the end of the project we expect to have a representative database for each of the languages under study and be able to work toward the long term goal of developing heuristics for the identification the type of abstract object referred to by a given nominal, with clear applications to automatic translation of nominal and anaphoric expressions. Our goals reflect the conviction on the part of the team members that 1) the integration of theoretical and practical application work in the area of computational discourse should serve as a model for how progress will best be made in computational semantics more generally, and 2) progress in linguistic theory depends on integrating large scale empirical work with the introspective methodology used in generative linguistics. Thus, a meta-objective of the project is to promote these two methodologies, both within the theoretical and computational linguistics communities, via the presentation of the results of our research, and via the training we provide to the young researchers participating in the project. Our group is made up of almost exclusively young researchers, for whom interdisciplinary research is of particular importance and impact. In fact, we plan for the younger researchers to participate in the most extended visits.

Our collaboration will be organized along an empirical dimension (one subgroup concentrating on the analysis of the reference of nominalizations, one on the analysis of anaphoric reference) and along a methodological one (theoretical, computational), dividing the team into interacting subgroups. Each group has members from each of the institutions.

Nominalization group: Katz, McNally, Alsina
Anaphora group: Bosch, Fontana, Rozario, Umbach,
Computational: Evert, Schrader, Katz, Boleda, Colminas

While maintaining strengths in both theoretical and computational research, the Barcelona group and the Osnabrücker group complement each other. The Barcelona group has more syntax-oriented researchers and particular expertise in the treatment of nominalization and nominals (and of course additional expertise for the Spanish and Catalan language), while the Osnabrücker group has particular strength in semantics and general computational-linguistic (and German expertise). Both groups are oriented toward addressing theoretical problems with computational means, however, and there has been significant ad-hoc collaboration between the two groups already.
Plan of the Project

To carry out our research we plan to combine:

- group meetings in various configurations, involving one or more subgroups, with the purpose of sharing results, refining the research objectives and the specific workplan necessary to attain them.
- individual visits by researchers from each lab, both to present research results and to acquire computational and research techniques from the other labs.

The individual visits will be especially useful for graduate students. We estimate that quarterly individual visits of 3-4 days of intensive work will provide important opportunities for collaborating researchers to make significant advances in their projects. These visits will be supplemented by e-mail, telephone contacts, and individual work by each team member at his/her home institution.

The project consists of three phases over the 2 year grant period.

Phase I (1-9/2006) will begin with a general meeting of the team. The goals of this meeting will be to provide team members with the opportunity to present their research in progress on the analysis of the phenomena central to the theme of the project (nominalizations, anaphora); to familiarize the members of the computational subgroups with these analytical problems so that they might suggest appropriate computational techniques for collecting relevant corpora to further the analysis; and to organize specific collaborations between the above-mentioned subgroups of researchers. The first phase will then continue in the form of individual and collaborative work along the lines planned, with a series of individual visits between the labs.

Phase 2 (10/2006-3/2007) will involve three meetings in subgroup format, one on nominalizations, one on anaphora, and one on corpus collecting/annotation. The first two (end of 2006) will involve members of each of the analytical subgroups and members from the related computational group (e.g. people working on the analysis of nominalizations/anaphora and those developing corpora for the analysis of nominalization/anaphora). The goal will be to obtain an inventory of the abstract entities necessary for the analysis of both phenomena. By the end of 2006 we expect to have a preliminary ontology which can serve as a basis both for further analytical research and for annotation. The third meeting (early 2007) will involve the whole computational group and two representatives of the analytical groups, to pool methods and results on the computational side and to prepare the format for annotating the corpus.

By Phase 3 (4-12/2007) members of the different labs will have had a full opportunity to share expertise and problems, and the first phase of corpus research will have expanded the relevant data base for the study of the analytical problems involved. This phase can thus be expected to be a very productive one, with drafts of papers being written, further quantitative corpus analysis being carried out and a multilingual annotated corpus being produced. This will again involve various individual visits between the three labs and will culminate in a final joint workshop with our Spanish colleagues in Barcelona (11/2007) at which the results of the collaboration will be presented and plans made for applying for future funding.

At the completion of the project we plan deliver a number of concrete results. First and foremost is an ontology that supports reference to abstract entities which supports the analysis of nominal and anaphoric semantics in texts of each of the languages under study. In addition, we will deliver a distributional analysis of the various linguistic forms under investigation in the languages under investigation and finally we will deliver a multi-lingual annotated corpus, with information about abstract reference annotated in the texts. We expect these results to be of considerable interest, both to theoretical linguists and natural language engineers.
Publications


