THIS PART CAN BE REVEALED TO THE APPLICANT

Project ID: P406/10/0434
Reviewer ID: 04025

QUESTIONNAIRE

1) Quality of the project proposal

1a) Originality, scientific importance, prospects of the project and expected benefits of the project for basic research

Characterize the purpose of the project; state in what way the project is relevant and promising; evaluate its competitiveness in the international context and compare its level with the current state of the art in the field:

The concept of the project is simple yet highly innovative and original: syntactically annotated corpus for the masses. Existing syntactic corpora for various languages, including the PDT Czech corpus, suffer from high theory-dependence. In case of PDT this is particularly acute, as it is based on a local dependency theory, not widely known to scholars outside of the Czech Republic. Hence, the need for a maximally theory-independent syntactic corpus, or treebank, of Czech is clear.

The project is scientifically important both for linguistic research and for natural language processing. For linguistics, the design of a formal syntactic representation corresponding to the traditional "school" grammar is bound to reveal both the strengths and the weaknesses of the pedagogical approach to syntax, and possibly lead to modifications of school curricula. The need to represent the whole variety of grammatical constructions, not just textbook examples, should spurn much linguistic research into the grammar of Czech. For the natural language processing, the project promises to contribute to the research on the combination of morphosyntactic analysis and stochastic parsing, and on error analysis leading to better parsing results.

Given its "second-order" approach to syntactic annotation, where various views of the underlying data are presented to the user based on his or her linguistic sophistication, the project is unique in the international context and should result in many conference presentation.

1b) Preparation of the project proposal, targets of the work and proposed deliverables

Evaluate the overall level of preparation of the proposal and the originality of the selected approaches to achieve the project's targets; evaluate planned deliverables (evaluate whether the targets set in the project correspond to the declared purpose of the project and how demanding they are):

The proposed deliverables are relevant to the project and reasonably ambitious. The only doubt concerns the amount of data that is be annotated automatically: it's a 1 million "experimental corpus". Given that these results will be manually corrected (p.8 of the proposal), this is a large amount, commensurate with the proposed budget. On the other hand, the proposal states on p.3: "We plan to provide syntactic annotation for the Czech National Corpus", suggesting the automatic annotation of the whole 500-million-word corpus. I would propose to follow this statement and annotate the whole CNC. Moreover, perhaps it would be beneficial to manually correct not an unspecified "experimental" corpus, but the 1.5-million corpus mentioned on p.3 that has already been annotated manually within PDT. This way the two annotation schemes could be accessed and compared automatically, possibly leading to interesting theoretical linguistic observations, and also this would be a way to train and better evaluate a converter from Functional Generative Description representations to the schema devised in the project.

1c) Concept, methodology and timeline

Evaluate whether the concept of the proposed work and methodology are clearly defined and the degree to which they are elaborated is correct; evaluate the proposed project duration in relation to its targets and the scientific importance of the project; evaluate the timeline of the project in relation to its feasibility:

The aims of the project are clearly defined and they justify the duration and the budget of the project. The project is ambitious but, given the team's previous rich experience in morphosyntactic analysis of Czech, in syntactic formalisms and in corpus linguistics, the project is feasible.

Quality of the project proposal is possible to evaluate as:

- [X] **excellent**
- [ ] **very good**
- [ ] **good**
- [ ] **satisfactory**
- [ ] **weak**

2) The applicant and his publication level and necessary facilities

Characterize the scientific level of the applicant and the team of workers in terms of their scientific results, number of publications (taking into account the age of the applicant), their quality and rating. State your opinion on the working capacity and facilities of the workplace:

The team is uniquely predisposed for carrying out the work described in the project. Various members of the team have rich experience with various linguistic formalisms, necessary for designing a formalism corresponding to the traditional "school" approach to syntax. Moreover, they have a long standing research record in morphosyntactic analysis and the morphosyntax-syntax interface. Particular team members have worked also on other issues important for the realisation of the project, such as collocations and valence dictionaries. The addition of Jiri Hana, a young but already established researcher, to the project's team, as planned on p.10, will further strengthen the human resources devoted to the project.

The head of the team is a prolific author: the project proposal mentions as many as 3 monographs since 2004, as well as some articles in proceedings and collections. The impact of these publications is limited, but that is unavoidable given the subject matter and the language of some of these publications (Czech).

The qualification of the applicant and the team of workers, their publication level and necessary facilities can be rated as:

X	**excellent**		**very good**		**good**		**satisfactory**		**weak**

3) Appropriateness and justification of the financial costs (Not necessary to evaluate)

X	**YES**		**NO**

Please state which requirements you consider unjustified:

OVERALL COMMENTARY ON THE PROJECT PROPOSAL

Please write your overall comments:

a) Strengths of the project proposal:

Originality of the concept "treebank for the masses". Expected research outcome, potentially important for theoretical linguistics, pedagogical linguistics, and natural language processing. Rich relevant experience of the team.

b) Weaknesses of the project proposal:

Limited amount of syntactically annotated data at the end of the project, but this is perhaps justified by time or budget constraints. Also, once the tools are produced within the project, they could be used to annotate the whole CNC after the project, with little additional overhead.

c) General comments:

Just a the Czech National Corpus turned out to be well-known and influential for the designers of other -- not only Slavic -- corpora, I expect the results of this project to reach well outside the Czech corpus community.