**Excellent**

**Fidler**

**Needle in Haystack Method for Czech corpus**

This project aims to implement an innovative methodology for finding topicality of targeted texts and to develop software that allows comparison between multiple texts. The initial effort will be to create a statistical analysis of word occurrence in an extremely large and well balanced corpus. This is the reference corpus (RefC). Indeed, it is the well balanced nature of the original corpus that is the foundation of project success. The Czech National Corpus meets this standard; for replication to other languages the absence of comparable corpora would be a limiting factor.

Subsequent to the development of the RefC, new texts will be tokenized and analyzed for unexpected frequencies of particular word forms. It is these unexpected frequencies that will the user to potential topics in the targeted texts as the T(arget)-texts can be evaluated against the reference corpus or against each other. The RefC can be divided in several ways including chronologically. This will facilitate not only for longitudinal studies of shifting word frequencies but for the creation of a RefC appropriate for a particular time period.

The widespread linguistic patterns will be stored in a Reference Corpus (RefC). In comparison to this reference key words (KWs) are words that occur in a text "more frequently than expected" given the RefC. The RefC will be based on the Czech National Corpus and the test cases to proof-of-concept are the New Years addresses of Gustav Husak from 1975 to 1989.

For this approach to work the RefC must be built from a large representative corpus, a characteristic that seems to apply to the Czech National Corpus although it is worth noting that the CNC does not seem to include transcription of oral speech. The T-texts are of a single genre and, ostensibly, author (though the effect of speech writers on the Husak corpus is not considered). The T-texts, as part of socialist diatribes (my word) do offer a chance for showing the utility of statistical KW analysis to show subtle changes in a somewhat standard matrix of politicized speech. The PI suggests that the Needle in Haystack Method (NHM) has "a high potential to detect subtle nuances" in the T-texts for this pilot study.

The description of the NHM software in the appendices is a clear presentation of a program that not only combines key word analysis (including relations among key words) with concordancing but allows for the comparison and analysis of multiple texts, which is key to the goals of this project. Key word is, in the context of NHM, defined statistically, as words that occur in a text more frequently than one would expect from similar analyses of a reference corpus. Users can eliminate certain words from statistical analysis and can "ignore case". In regard to the latter it is not clear whether this refers to orthographic case (upper and lower) or to nominal cases (and perhaps more generally, declension). It seems, given the absence of further discussion, that it refers to orthographic not grammatical case.

It is interesting that the NHM explores word and not lemma frequencies. I was struck by this fact and wondered how difficult it would be to build in declension to the search engine/statistical analysis. It would seem possible to build a "behind-the-scenes" functionality to neutralize (or not neutralize, depending on user choice) declension of Czech nouns. One reason is perhaps that the project is designed to require no "knowledge of corpus-/computational linguistics (no complicated automatic pre-processing of T-txts, no required installation of a local computer)".

Apparently, then the choice not to use either a parser for lemmatization or to build a back-end tool for partial lemmatization (e.g., of nouns) was a conscious decision based on a desire to keep NHM as simple as possible. On p. 7 they do mention that Czech is highly inflected and indeed state that this is a positive feature for their "word-form" approach as the forms "contain valuable information about varying degrees to which individuals play an active role in an event (e.g., actor, patient, recipient, experiencer, possessor, partner)". This is true but it is also true that parsing out the grammatical information would better allow such role analysis as the grammatical function would be found with lemmas and be searchable across word forms.

The PI also notes that it is meant not as a diachronic study of grammar (or semantics) but rather "it aims to investigate textual indicators of societal-cultural shifts as well as the trajectory in which the society might be headed" (p. 7). They might have also added that it is not a study of diachronic lexicosemantics (cf. Geeraerts's book on this topic) as it is the word form in isolation that is statistically analyzed and not any potential shift in meaning. Indeed, it is this potential for shift in meaning that could create a problem/contamination of the frequency of occurrence over time. I am not sure if there is way to control for semantics (polysemy, diachronic shift) but it is a concern that might have been addressed.

Finally, the ability to compare texts goes beyond what other tools offer: to find patterns within a single corpus. The NHM is based on a statistical analysis of saliency not user defined searches for words the user wants to find. NHM is more data-driven (based on the RefC ) and it can contextualize texts in time defined periods of language use.

In terms of replicability, the NHM would seem to require the existence of a large and balanced corpus of materials to construct a RefC. It is not clear how many languages have corpora that meet this requirement. Another question involves the utility of a methodology of statistical analysis of word forms for languages with highly complex morphology.

**Overview**

Intellectual significance: This is an extremely interesting proposal that has implications for textual analysis beyond the material to be development in the present project.

Impact on research and technology: The need for a large and balanced corpus to create the RefC will limit the extension of the NHM to other languages. As the PI notes the problem with corpus approaches to measuring the statistics of word occurrences is that of using a balanced corpus. Thus Google ngram uses digital material on the Internet. Nevertheless, the ability to obtain significant results from a methodology that involves no parsing or semantic analysis could be very important for future attempts in a similar vein, particularly as large corpora are developed.

Innovation: The project is innovative in concept, design, and implementation.

Proposal development: The proposal is a model of clear and concise writing, with an excellentpresentation of the product and of the "environmental scan".

Feasibility: The prototypes developed to date suggest a very high level of probability that the project will be successfully completed.

Qualifications: All members of the team are highly qualified.