In coordinated efforts at Charles University (Prague, The Czech Republic) and the University of Pennsylvania (Philadelphia, USA), we propose to advance the theory and practice of linguistic annotation of text. At these institutions, there is a rich tradition in computational linguistics and in the creation of annotated corpora, and several active on-going annotation projects, in which there is already collaboration between the sites. We are proposing a new project that focuses specifically on the content and process of linguistic annotation itself, in an effort to solve some long-standing problems that are blocking progress in the basic science and technology research that these corpora support.

Over the past decade, the science and technology of human speech and language have been revolutionized by the availability of large amounts of linguistic data in digital form. Beyond the fundamental requirement of raw speech recordings and digital text, a crucial role has been played by a variety of descriptive and analytic notations applied to this basic data, including especially the annotation of syntactic structure (“part of speech tagging” and “treebanking”), shallow semantic structure (“sense tagging,” “entity tagging,” “co-reference tagging,” “proposition banking” and the like), as well as various aspects of discourse structure and spoken performance.

It is hardly an exaggeration to say that the field of computational linguistics has redefined itself as the problem of modeling such databases of linguistic annotation. This is equally true for the development and evaluation of new methods, such as new parsing techniques, and for the application of existing methods to new problems, such as the analysis of new languages. In addition, allied fields are being influenced more and more strongly by the concepts and the methods of corpus-based computational linguistics, for instance models of syntactic processing in psycholinguistics and studies of syntactic change in historical linguistics.

These changes have been driven by the success of linguistic analysis (and synthesis) algorithms based on empirically-trained statistical models. While these algorithms are by no means perfect, they usually perform much better than the systems based on hand-crafted rules that they have largely replaced; and they are portable, in the sense that they can in principle adapt to new training data – from a new language or a new problem domain – with little or no new human labor.

However, there are several dark clouds in this otherwise bright scientific and technical sky. The available annotated corpora are too small and do not cover several text genres of practical importance. The existing annotation standards ignore many crucial phenomena and mis-represent some others; only a handful of languages have any annotated corpora available at all; porting annotation techniques to new languages and even new text genres is
a painfully time-consuming process, typically requiring more than a year before production begins in earnest; training new annotators for existing languages and genres takes months at best; high-quality (i.e. inter-subjectively consistent) linguistic annotation is slow and therefore expensive; and new linguistic annotation is carried out on a significant scale only in a bare handful of places.

As a result, the current rate of creation of new linguistic annotation is distressingly small, largely because the unit labor costs are very high, and also because the needed expertise is not widely distributed and is expensive to teach. The start-up costs (and delays) for new languages and even new genres of text are remarkably high – that is, skilled computational linguists, annotation managers and annotators must work together for many months before tasks are defined, tools are ported, annotators are trained and results begin to emerge from the pipeline. As a result, while current algorithms can be ported quickly, easily and cheaply to new languages and new domains, the necessary training data is not at all quick, easy or cheap to create. In addition, ideas for new types of analysis, for instance in the area of semantics, face an uncertain future because of the high costs and long time-scales for creation of adequate annotated data for experimentation.

While there are promising ideas for substituting unsupervised learning based on very large raw (unannotated) corpora, most researchers in computational linguistics and related fields believe that supervised (or semi-supervised) learning, based on annotated corpora, will continue to play a crucial role for some time to come. As a result, the speed and cost of annotation remains a significant roadblock to progress in the science and technology of speech and language.

We propose to improve this situation by treating linguistic annotation as a research problem in itself. For the purposes of this project, we will focus on the annotation of text (including transcripts of spoken language as well as material that was originally produced in written form). We believe that such an effort will make significant improvements in many aspects of the linguistic annotation process. Specifically, we aim at coordinated improvements in coverage, in productivity, and in portability. These improvements will also make training easier, and will increase both the number of sites at which such annotation is done, and also the overall number of competent annotators.

Quantitatively, we would like to improve typical productivity by a factor of ten, and to make a similar increase in the overall number of competent annotators, so that the volume of production can be increased by two orders of magnitude. We also hope to reduce the time to port treebanking and similar annotation schemes to a new language by an order of magnitude. To give some specific numbers, we believe that it should be possible to increase typical treebanking production rates from roughly 150 words per labor hour to 1500 words per hour, and to decrease the start-up time for treebanking in
a new language (for entering stable production mode) from about 400 days to about 40.

Qualitatively, our goals include fixing the areas where existing morphological and syntactic annotation standards are missing or wrong; extending existing semantic annotation standards towards integrated coverage of word sense resolution, predicate-argument structure, and key aspects of reference and scope; and combining these with existing models of pragmatic annotation (such as functional sentence perspective, discourse and rhetorical structure). In terms of language coverage, we aim to produce consistent tools and annotation specifications, for a wide enough range of languages to make porting to most new languages significantly easier than it has been up to now. Finally, we will be working in two superficially different analytic frameworks – roughly those of the Penn and Prague Treebanks – with methods for inter-translation.

We hope to accomplish these ambitious goals efficiently and cheaply because the cooperating institutions, in addition to considerable past experience in annotation of this type, have several on-going annotation projects that are independently funded. For example, at Penn there are on-going projects on annotation of biomedical texts, texts in Chinese and Arabic, and texts throughout the last millennium of the history of English. These projects are produce annotated corpora and/or to do research based on such corpora, not to work on the problem of how to do such annotation better and faster. However, if we add a component of research on the annotation methods themselves, the existing projects provide testbeds for studying the annotation process, pinpointing problems, and evaluating new methods.

Based on extensive past experience in several kinds of linguistic annotation, we have specific ideas about how to approach our goals: modifications and extensions in annotation standards; changes in workflow, tool design and training; new kinds of annotation and new relations among existing annotation outputs. However, the crux of our proposal is to focus the attention of a group of faculty, staff and students on the problems, at two institutions where much of the world’s linguistic annotation is being done, through a coordinated program of individual research, systematic implementation and testing, with yearly workshops for sharing results and joint exploration of new ideas. These workshops will also serve as a mechanism for promulgating, to a larger community, the skills and tools involved in the efficient creation of large linguistically-annotated databases.
Insititutional Background

The University of Pennsylvania (USA) and the Charles University (Czech Republic) have longstanding experience with rich, extensive and technologically advanced linguistic annotation. Over the past decade, researchers at both sites have created and published many linguistic resources that have been used in the development of speech and language technology, and also in research on language structure, history and use.

Penn and Charles University have a long history of informal cooperation, ranging from theoretical computational linguistics to close cooperation on Arabic annotation. These relationships involve several different academic departments and centers on each side, and will serve as the core of the proposed educational cooperation.

At Penn, the institutions involved include the departments of Linguistics and Computer and Information Science (CIS), the Linguistic Data Consortium (LDC), and the Institute for Research in Cognitive Science (IRCS). IRCS will be the lead organization on the Penn side, and the Philadelphia workshops will be integrated administratively with the yearly IRCS/CCN Workshops in Cognitive Science and Cognitive Neuroscience, which have been held for the past several years.

The Penn researchers involved include those responsible for the design, implementation and publication of treebanks in English, Chinese, Korean and Arabic; a historical treebank of Middle English; an English “proposition bank”; the discourse and disfluency annotation of Switchboard; several years of content annotation associated with the MUC and ACE projects; application of entity tagging, part-of-speech tagging and treebanking to biomedical text; and many other relevant examples of annotated corpora.

At Charles University, the lead organization will be the Institute of Formal and Applied Linguistics (Faculty of Mathematics and Physics), with the involvement on as-needed basis of the Center of Computational Linguistics (Faculty of Mathematics and Physics) and the Institute of the Near East (Faculty of Arts).

Computational Linguistics has a 45-year long tradition at Charles University, supported by previous research of the Prague Linguistic Circle members dating to the early 1920s. A project for building a large electronic Czech National Corpus was started once the political and technical environment became suitable in the early 1990s, with complex linguistic annotation projects starting as early as 1995. The Czech language is in this respect one of the best served languages of the world now, with an available national corpus of half a billion words of text, more than a million-word corpus with rich morphological and syntactical annotation, and annotation software support and many NLP tools developed at Charles University.
The Czech participants in this project are those who have organized the annotation efforts at Charles University, from linguistic annotation specification to software to the manual annotation and its organization.

**Problems and solutions**

We will divide this discussion into six headings: annotator productivity; annotator training; annotation standards; integration of annotation types; start-up time for new languages and genres; and introductory teaching materials.

**Annotator productivity.** There are now two basic methods of annotation in common use: human end-to-end annotation, and machine annotation followed by human correction (often called “semi-automatic” or “machine-aided” annotation). Research (as well as folklore) indicates that the “semi-automatic” methods are significantly faster, and also more consistent, though at the cost of some machine-created bias in the errors.

However, the resulting methods remain very labor-intensive. Overall productivity measures (total output divided by total labor hours) yield numbers like 150 words per hour for syntactic treebanking, or worse. This is roughly 60 times slower than the rate for careful reading out loud, and somewhere between 100 and 200 times the typical rate for careful silent reading. If the reader does not find this excessive, (s)he should read this sentence out loud 60 times, and judge at what point its structure has been adequately understood.

Careful examination of the process suggests that the problem is not mainly due to poor user interface design, inadequate management oversight, or any other simple mechanical or organizational factors. That is not to say that improvements in tools, motivation, workflow monitoring and so on are impossible. However, the basic problems seem to be intrinsic in the tasks as they are currently defined.

Key issues include very long learning curves for new annotators; the need for multiple passes to ensure reasonable quality; the additional overhead of discussions and meetings to achieve and maintain analytic consensus; the very high cognitive complexity and cognitive stress of many of the subtasks involved; and large amounts of time spent on a small fraction of especially difficult cases.

Some directions for solutions are clear.

First, it makes sense to interleave human and machine processing in a more coordinated way. For example, current parsers are bad at coordination and PP attachment; but with the right interface, a human reader can sketch the scope of conjunctions and the attachment point of PPs in just a few times
reading time, i.e. at rates of several thousand words per hour. Once a modest amount of information of this type has been provided, current parsers will make very few mistakes. It’s likely that they will make no mistakes at all on many sentences, so that correction can become much faster. We should contemplate more elaborately layered approaches, in which several very efficient human passes (each working at several thousand words per hour) are interspersed with appropriate machine processing.

Second, it will obviously help to apply some sensible “cognitive engineering” both to the visual representation of linguistic structure and to the judgments that we ask annotators to make in registering it. There are many aspects of current annotation practice that are hard to learn, and hard to read and understand even after they have been learned, but where the information being supplied can usually be described in a way that is simple and intuitive. We should evaluate ideas of many kinds in this connection, including ideas not only from Penn-style labelled bracketings and Prague-style lexical dependencies, but also TAGs, categorial grammars and construction-based approaches, since each of these has certain phenomena for which it provides an especially intuitive account. In a properly designed annotation task, it’s likely that each human pass will be asked to supply just one or at most a small number of types of information, and so we can focus on the best way to indicate the choices and register human judgments for each type of information.

Third, we need to experiment with various approaches to “quick annotation.” In the case of speech transcription, careful transcription of conversation takes 50 times speech time even for highly practiced transcribers. However, we were able to reduce the needed labor to 5 time speech time by several techniques, of which the most important was the simple instruction to skip phrases that are too difficult to transcribe quickly (usually because of complex disfluencies or unclear speech), or contain other problems that would cause a delay (such as a proper name or rare word whose spelling needs to be checked). On 20 hours of conversation, this type of “quick” (though incomplete) transcription was tested against an existing full, careful transcription, from the point of view of its value in training a speech recognition system. The recognizer trained on the “quick” transcriptions performed slightly (though not significantly) better than the recognizer trained on the full transcriptions; of course, for the same amount of money, the “quick” method could create 10 times more data and thus a much better outcome.

Finally, we should experiment with various sorts of “active learning” or other stratified sampling, where the materials to be annotated are selected so as to be enriched in phenomena of interest, or where a much larger volume of material is annotated more sparsely, along with a smaller volume of more densely annotated material.
**Annotator training: methods and materials.** We need a general framework and toolkit for self-paced training methods, along with the content for training in key areas such as treebanking and “prop banking” in English. To the extent that we are successful in cognitive re-engineering of annotation tasks and in creating tools that can facilitate layered application of human and machine processing, providing self-training materials will become easier.

**Annotation standards.** A number of things are missing or wrong in the current annotation standards. For example, the Penn Treebank standard for syntactic annotation does not attempt to assign any structure at all to the pre-head portion of noun phrases. This omission is disastrous for technical writing, where half of all the relevant “facts” for an information extraction task may be tied up in relations among pre-head nominal subconstituents such as “adrenaline-induced NK cell activity enhancement.” The Penn Treebank standard treats hyphenated sequences as single words, which produces a hash of examples like

(NP (DT the) (NNP Charles) (NNP MacArthur-Helen) (NNP Hayes) (NN saga) )

and the much commoner analogous cases in technical writing. The MUC and ACE entity tagging and “co-reference” standards prescribe some philosophically-odd treatments of indefinite noun phrases, especially in embedded structures, that don’t generalize well to biomedical texts.

One of the key difficulties in applying annotation standards in new domains is that omissions (like the lack of pre-nominal structure) or mistakes (like the treatment of hyphenated sequences) that are innocuous in one genre (like Switchboard conversations or Wall Street Journal news stories) may turn into show-stopping problems in another domain. There are analogous effects in porting to new languages.

While “universal annotation standards” are a utopian goal, a sensible process of improvement would result in standards that are both more useful and easier to apply in new cases.

**Annotation integration.** There is a general need to provide better integration of different types of annotation. The development of stand-off annotation in general, and of the “annotation graph” formalism in particular, makes it possible for a wide range of different types of annotation to co-exist without mutual interference of the type that in-line annotation mark-up creates. However, this is not enough. There are many cases in which different sorts of commonly-done annotation create mutual logical dependencies that are not enforced, and more important are not taken advantage of. For example, so-called “entity tagging” (in which a certain typically nominal word or word sequence is marked as a reference to an entity of a certain type, such as a person or a chemical compound) has never been systematically connected with part-of-speech tagging or treebanking.
As types and subtypes of annotation proliferate and become richer, we need a general way to think about the dependencies among them. These dependencies can then be used to allow progress in one type of annotation to help in other areas. For example, bottom-up “chemical compound taggers” have achieved excellent performance on biomedical text, where the correct delimitation of multi-word compound names is a significant problem for general-purpose parsers – and also for human treebankers.

**Start-up time for new languages and genres.** In the end, so far, new language and genres always wind up looking surprisingly similar to old ones. It is mostly coming to understand how to establish the similarity that takes so long. In addition, it must be noted that even after annotation specifications, tools and cultural practices are well established, training a new annotator from a standing start may take as long as six months for some kinds of annotation (such as treebanking). So the fact that treebanking for a new language takes something like a year to get going at a production level is partly because of this more general long training time, which is compounded with difficulties in creating new specifications, adapting tools, etc.

To some extent, porting to new languages will thus be speeded up by anything we can do to speed up the training of new annotators in general. The development of more general tools and sounder, more rational annotation standards will also help. However, this is surely the most difficult problem that we face.

**Introductory teaching materials.** In principle, linguistic annotation is a good way to teach linguistic analysis; and the resulting skills are exactly those that will enable a student to think analytically about specific problems in language structure and use. The “cognitive engineering” required to make the descriptive categories accessible to annotators is likely to make the descriptions accessible to students as well. The semi-automatic mixed-initiative tools that will make the annotators’ job easier could also help guide and correct students. Finally, the self-paced training materials needed to make it possible to enroll and train new annotators efficiently are also just the sort of thing that students need to learn techniques of linguistic analysis.

In the context of this project, the main motivation for helping to produce introductory teaching materials is that introductory courses are a good recruiting ground for an annotation labor force. Students learn what annotation is like, and can evaluate how much they enjoy such work by comparison to other alternatives. Project managers can recruit workers who have already learned the basics.
The role of yearly workshops

A crucial aspect of the proposal is a series of yearly Linguistic Annotation Workshops, to be organized jointly and held alternately in Philadelphia and in Prague. The first workshop will take place in Philadelphia in late June of 2004.

These workshops will have three coordinated components: a research component, an information exchange component, and an educational component. Each year’s workshop will have a small number of research themes that will receive special attention in all three dimensions, though other aspects of annotation will also be treated. Examples of possible themes include the problems of porting annotation techniques to a new language or domain; the relationship of monolingual and bilingual lexicons to multilingual parallel annotated data; integration of syntactic and semantic annotation; achieving qualitative productivity improvements in large-scale annotation.

For the research component, the workshops will enable a mix of senior and junior researchers and students from both participating institutions, as well as some outside participants, to work intensively and collaboratively for a few weeks on a few key problems of linguistic annotation. This intense, focused collaboration will occur in the context of longer-term background activities to prepare for the workshops and to follow up on their results. These intensive discussions will also make it possible to reach a consensus on new aspects of annotation content.

For the information exchange component, we will have the usual kind of reports on research results and discussions of future plans. However, the format will permit much more detailed presentation and discussion than can be achieved in presentations at existing conferences, and will involve a larger number of participants in on-going annotation projects than normally travel to any particular conference.

For the educational component, there will be lectures, laboratory exercises, and practical projects, with a special focus on the current year’s research themes. Many of the students involved will also take part in on-going annotation projects outside of the workshop structure. Lectures for all workshops will be prepared jointly by the senior staff of the two institutions, 3-4 of whom will travel each summer from Prague to Philadelphia or vice versa.

We envision the teaching of annotation schemes developed at both institutions (for syntax, semantics, discourse structure, entity and relation tagging, etc.), with examples from at least four languages in which annotation projects are underway at the two institutions (English, Czech, Arabic and Chinese). Technical aspects of annotation (markup schemes and languages, tools for annotation, quality control and workflow management, taggers/parsers, etc.) will be covered in detail. We will also discuss
organizational aspects, such as developing annotation guidelines, recruiting, training and managing annotation teams, and dealing with intellectual property issues in planning for publication.

The core participants in the research and information-exchange components of the workshops will be faculty, staff and students from the two participating institutions. Others will be selected as individual interested participants, or as representatives of groups at other institutions (especially in other countries) where similar projects exist or may be started.

For the educational component of the workshops, students will be selected by a nationwide search in the U.S., and by a similar application process on the Czech side. About 10-15 students are envisioned for this aspect of each workshop, in order to ensure intensive mentoring by participating faculty. Student travel and accommodation will be provided.

**Broader impact**

The primary impact of this work will be to enable the creation and publication of significantly larger, higher-quality annotated text corpora, at a significantly lower cost of production. We aim for an order-of-magnitude change in unit costs, and another order-of-magnitude change in the world-wide number and size of successful annotation projects, so that the rate of creation of such data can increase by two orders of magnitude. Since modeling such data has become the central activity of computational linguistics and related areas, any success towards these goals will have a significant impact on the work of thousands of researchers world-wide.

Additional related activitives will further promote the development and use of linguistic annotation schemes in the computational linguistics community around the world. For example, tutorials will be prepared as a short version of the lectures given at the workshops, and submitted to various conferences with which tutorials are usually held, such as the Association for Computation Linguistics. The first such occasion will be at the ACL04 in Barcelona, Spain.

In addition, we will propose special sessions or satellite workshops on each year's key research themes, to be held at meetings such as LSA, ACL, ICSLP and so on, and involving research presentations by workshop participants as well as others.

Finally, we hope that our development of basic teaching materials, which in the context of this proposal is simply a method for recruiting and screening potential annotation workers, may contribute to others’ efforts to improve education in those neglected foundations of a liberal education, grammar and rhetoric.