DEVELOPING LINGUISTIC RESOURCES WITH THE ANC
REPORT ON THE NSF-FUNDED WORKSHOP

Nany Ide, Vassar College

Christiane Fellbaum, Princeton University

1 Introduction

An NSF-funded workshop on Developing Linguistic Resources with the American National Corpus (ANC) was held October 29-30, 2006, in New York City. The workshop brought together representatives from academia, industry, and government agencies who are involved in developing and using linguistic annotations of language data, to consider the ways in which the ANC can best be developed to serve the needs of computational linguistics research. Specifically, the workshop addressed the following topics:

- the direction and development of the ANC, and the kinds and amount of linguistic annotation of the ANC that would best serve the community's needs;
- collaborations and cooperation among annotators and developers of other linguistic resources that can be established in order to further enhance the ANC as well as other projects and resources;
- the needs of the CL research community for language data, annotations, and other linguistic resources, and the steps that should be taken to answer these needs.

The following persons participated in the workshop:

Collin Baker, International Computer Science Institute / University of California at Berkeley
Hans Boas, University of Texas, Austin
Branimir Boguraev, IBM
Nicoletta Calzolari, Istituto di Linguistica Computazionale, Pisa, Italy
Christopher Cieri, Linguistic Data Consortium / University of Pennsylvania
Christiane Fellbaum, Princeton University
Charles Fillmore, International Computer Science Institute / University of California at Berkeley
Sanda Harabagiu, University of Texas, Dallas
Rebecca Hwa, University of Pittsburgh
Nancy Ide, Vassar College
Tanya Korelsky, National Science Foundation
Judith Klavans, Center for the Advanced Study of Language / University of Maryland
Erin McKean, Oxford University Press
Adam Meyers, New York University
Joseph Olive, DARPA

Martha Palmer, University of Colorado
Becky Passonneau, Columbia University
James Pustejovsky, Brandeis University
Janyce Wiebe, University of Pittsburgh

This report summarizes the workshop discussions and outlines the suggestions for future activity determined by the workshop participants. The report is available on the ANC website: http://AmericanNationalCorpus.org/nsf-workshop-2006. Comment and response from the research community is invited.

2 Background and Motivation

2.1 American National Corpus Project

The following was provided to the workshop participants prior to the meeting, and provides an outline of the activities the ANC, together with the FrameNet and WordNet projects, planned to undertake in the future. The goal of the workshop was to solicit input from the community on the proposed plan, and to modify it, if necessary, to best serve the community's needs. The accompanying Powerpoint presentation is available at http://AmericanNationalCorpus.org/nsf-workshop-2006.

The American National Corpus (ANC) Project is creating a large corpus of contemporary written and spoken American English data representative of a wide range of language styles and types, which will be richly annotated with linguistic information so as to serve as a resource for linguistics and computational linguistics research. To date, the ANC has released 22 million words of varied written and spoken data, available through the Linguistic Data Consortium (isbn:1-58563-369-0), annotated with two different part of speech/lemma schemes as well as noun chunks and verb chunks. When completed, the corpus will consist of a 100 million word core, comparable to the British National Corpus in genre distribution, together with a "varied" component including newer genres such as blogs, discussion lists, email, etc. As the only large corpus representing contemporary English usage and including a representative range of genres, the ANC's value for language processing research is evident.

The centerpiece of the ANC will be a 10 million gold standard sub-corpus consisting of a representative sample of written and spoken language types, containing hand-validated annotations for word and sentence boundaries, part of speech, syntax, and named entities (people, locations, organizations). We plan to manually annotate/correct a portion of the gold standard corpus for WordNet senses and FrameNet frames, to serve as a testing ground for harmonization of these two major resources and provide means to bootstrap more accurate automatic annotations for both WordNet senses and FrameNet frames in the remainder of the ANC.

The cost of manual annotation for a wide range of linguistic phenomena (e.g., syntax, co-reference, semantic information of various types, discourse structure, etc.) for the entire ANC is prohibitive, and therefore the ANC Project plans to annotate the entire 100+ million words using available automatic annotation tools, providing in some cases multiple annotations of the same type (e.g. syntactic analyses produced by several different parsers). These annotations can serve as a basis for comparison and merging of

different annotation tools' output and can be used to improve their accuracy through techniques such as machine learning.

Perhaps most importantly for NLP and linguistic research, the ANC project has implemented a scheme for representing the ANC and its annotations that answers one of the field's urgent needs, that is, means for individual annotations to cohabit with one another and/or be layered over primary data that allows them to be easily merged and manipulated. The perennial problem for language processing research is the fact that annotations produced at different sites are idiosyncratic and often demand considerable processing to render them usable with software for which they were not originally designed, or to combine annotations produced at different sites. For example, the *Wall Street Journal* corpus has been annotated by several groups for different phenomena: The Penn Treebank (PTB) II syntactic annotations are embedded in the data itself in a LISP-like format; PropBank and NomBank reference the data by sentence (tree) number and token number; and the Penn Discourse Treebank uses a complex addressing system to reference nodes in the PTB constituency trees. Using any one of these annotations demands that one's software can process their addressing mechanisms (which typically requires programming effort to adapt), and merging them is very far from a trivial task. The result is that individual projects and groups spend considerable time and effort to massage annotation information into a format they can use, again involving much duplication of effort and time spent on low-level tasks prior to addressing research questions.

The philosophy of the ANC scheme is maximum flexibility. This is accomplished by first rendering all annotations into a generic, feature structure-based format[1] and outputting each in a separate stand-off document linked either to the primary data (which is text-only and read-only) or to other annotations. Annotations in this format utilize the *original annotation labels*—i.e., we make no effort to provide annotations harmonized on linguistic grounds. Users then use a graphical interface to the ANC's freely-distributed ANCTool[2] to select the annotations they are interested in, and the tool produces a version of the corpus with the merged annotations in-line. The output format of the data and individual or merged annotations is controlled by the user, and can therefore be tailored to a particular software system or use. The current version of the ANC tool provides three built-in output options, including one XML format[3], suitable for input to the BNC's XAIRA system, and two non-XML formats for use with MonoConc Pro and WordSmith, respectively. This choice of options was dictated by the need to accommodate the large number of corpus linguists initially using the ANC. However, because the tool's underlying parser uses multiple implementations of the org.xml.sax.DocumentHandler interface (one for each output format), additional formats are trivially generated by implementing additional interfaces. For example, we have recently implemented options

---

[1] The ANC format is isomorphic to UIMA's Common Analysis Structure representation, which also uses feature structures.

[2] See http://AmericanNationalCorpus.org/tools.html

[3] For XML output, the user also chooses how to handle overlapping hierarchies from among several options.

to generate output in a form suitable for use with the WordNet project's annotation software, as well as a graph representation that serves as input to GraphViz[4] so that annotation structures can be visualized.

The ANC is distributed by the LDC. It is free (apart from a $75 processing fee) for research. Users sign two licenses, one for a "restricted" portion of the corpus, and one for an "open" portion. The open portion, which is freely re-distributable and re-usable, comprises about 80% of the 22 million words currently released.

2.2 Discussion Points

Prior to the workshop, the following discussion points were distributed to the participants.

ANNOTATION TYPES AND REPRESENTATION

1. Which annotation types (i.e., linguistic phenomena) are most urgently needed for CL research, and will be needed in the next 5-10 years?
2. The ANC allows multiple annotations of the same or similar types to co-exist:
    a. Is the availability of multiple annotations for the same phenomenon (e.g., multiple syntactic analyses) a valuable resource for CL research?
    b. For which linguistic phenomena would it be valuable to have annotations of the same or similar phenomena reflecting different theories or schemes (e.g., PropBank and FrameNet annotations)?
3. The ANC representation format allows annotations to reference not only the primary data, but also other annotations (e.g., a co-reference annotation for noun phrases). Is this kind of annotation layering desirable, and is it something that other annotation projects are doing or are capable of doing? If so, how can we encourage collaboration and avoid duplication of effort?
4. In the gold standard sub-corpus, only one annotation for each linguistic phenomenon can be hand-validated. Which syntactic analysis should be in the gold standard (constituency vs. dependency, which parser's output)? Are there preferences for other phenomena?

ANC USABILITY

1. Beyond the 100 million word core corpus, which will contain the same distribution of genres as the BNC, the ANC will include additional materials consisting of more recently developed genres such as blogs, email, chats, etc. as well as additional data for genres already represented in the core. Are there any genres that are of particular interest to the CL research community that we should focus on obtaining?

---

[4] http://www.graphviz.org/

2. Is there a particular output format(s) for merged versions of the corpus that would be most useful for CL research (e.g. LISP-like formats for Treebank style annotations, or other non-XML formats), that we should provide as an option for output of the merging tool?
3. Is there anything that you would like to see in the ANC, or any way for it to be organized or annotated, that is not already in the plan?

ANNOTATION STRATEGIES

1. We plan to annotate at least a portion of the ANC data (e.g. a representative sample of the gold standard sub-corpus) manually for WordNet senses and FrameNet frame elements, and eventually use the manual annotations to improve the performance of automatic sense and frame element taggers. Another approach is to annotate the entire ANC for these phenomena using possibly several different automatic annotation systems for each phenomenon, and manually correct some portion of the annotations and/or examine results from different systems to see if combinations of results can be used to correct at least some portion of the automatically-produced annotations. We would appreciate recommendations on the best way to go about this. Also:
    a. To what extent should we attempt to manually annotate the same words with WordNet senses and FrameNet frames? Should we attempt to annotate frame elements (to the extent possible) with WordNet senses as well?
    b. What steps can/should we take to harmonize these two resources in the course or as a result of annotation of the ANC?
    c. Are there particular improvements in WordNet and FrameNet that you feel we can/should we strive for when annotating the ANC?
2. We intend to use state-of-the-art methods for improving the performance of automatic annotation tools, by exploiting the manual annotation of the gold standard sub-corpus together with machine learning techniques, comparison/merging of annotations for the same phenomenon by different systems, etc. How can we best go about doing this, in terms of the phenomena and methods to begin with? Should we begin by comparing multiple annotations of the same type, or focus on one annotation and use strategies such as active learning to improve automatic annotation performance? Should the strategy differ for different phenomena?

OUTREACH/COLLABORATION ISSUES

1. The ANC is heavily used by corpus linguists and CL researchers in Asia, but it has not yet been used much by US researchers. How can we increase the visibility and use of the ANC within the US CL community?
2. One of our goals is to have annotations of the ANC produced and contributed by members of the CL research community, which we would then render into the stand-off format and distribute freely for use by the entire community. How can the ANC

best reach out to the CL research community to encourage either manual or automatic annotation of the ANC?

3. Are there collaborations with members of the CL research community that we could/should enter into to enhance the development and usefulness of the ANC?
4. The ANC has already collaborated closely with the GATE team at Sheffield University, whose software has been used to process all data and annotations, and the ANC can be generated in a GATE-compatible format. Are there other annotation systems we should accommodate, or other teams we should collaborate with?
5. How can we ensure continuous input on ANC development from the CL research community?

GENERAL ISSUES FOR THE DEVELOPMENT OF LINGUISTIC DATA AND ANNOTATIONS

In addition to soliciting input concerning the development of the ANC, we would like to open a general discussion on the ways in which the US computational linguistics research community can work together to enable greater consistency and inter-operability among linguistic resources, including both annotated data and supporting resources such as computational lexicons, frame banks, etc. At present, numerous annotation-related efforts are on-going in relative isolation or, at least, without much attention to the need for consistency and inter-operability. It is certainly premature to consider any given theory or approach as the "correct" one, and we can all benefit from comparison and collaboration to achieve our common goal of enhanced language processing capabilities. To that end, we ask you to consider the following questions:

1. What are the major obstacles to general reusability of annotated corpora and other linguistic resources to support NLP, and what steps can be taken to overcome them?
2. What would you recommend to funders of US CL research as the best way to move forward in the creation of corpora and other language resources so as to maximize the potential for progress? Please consider existing resources that should be enhanced, non-existent or relatively inaccessible resources that should be developed, ways to ensure accessibility of resources for research, ways to ensure reusability and consistency of resources, etc.

3 Discussion Summary

3.1 Annotation Types

There is an urgent need for annotation of many linguistic phenomena, including but not limited to word senses, syntax, multi-word expressions, discourse, opinion/point of view, event structure, causation, expressions mentioning named entities, and connections between discourse level and other phenomena.

Data should be annotated from different (possibly competing) theoretical approaches; for example syntactic annotation using phrase structure and dependency syntax, in order to support research that compares the various approaches to show both how they relate and which are more appropriate for a down stream use.

In general, annotation should be application-driven (e.g., discourse level annotations useful to Question Answering).

3.2 Representation

Selected access to annotations should be provided, so that users can choose to work with the annotations of interest to them. The means to accomplish this was generally referred to as "layering" of annotations using a stand-off approach, although there was no discussion of how to determine what the layers would consist of.

Mechanisms to combine (link) annotations of different types should be available (e.g., FrameNet and WordNet, TimeML and PropBank, etc.). The ANC approach of producing a graph from several different annotations, in which commonly annotated data spans are identified, was regarded as promising, especially for the work within the ULA project to merge PropBank, NomBank, and TimeML annotations.

3.3 ANC Usability

3.3.1 Data types

Annotated texts should come from many different genres; some topic-specific and some genre-specific. However, there is a serious lack of suitably diverse language data currently available.

Several new genres are increasingly important for language processing applications, blogs, wikis, social networking, and games. The ANC has the unique opportunity to capture these new communicative modes, including including "news talk", closed-caption text, email, newsgroups, blogs, chat, wiki, etc.

The question arose as to why another corpus (the ANC) should be created and more annotation be performed, that is, why are the data that are currently made available, for example via the LDC, not appropriate or sufficient for the uses targeted by the ANC. Chris Cieri, who represented the LDC at the workshop, provided several motivations for the ANC project. First, the ANC is a dynamic enterprise; the corpus is being built up continuously and in part from current data from emerging genres. In comparison, many LDC are used as benchmarks, reference corpora used to measure technology performance. To support this use, benchmark corpora need to remain, reliably, in their published form. The LDC corpora that are treated as dynamic, for example the so called Gigaword News Text corpora and the data supplied via the LDC online service, address only a small subset of the genres targeted by the ANC. Second, the LDC is by necessity opportunistic. Both the data donated to LDC and those created directly by the LDC are the results of – and thus address the linguistic genres needed by – the sponsoring programs. As a result, the ensemble of corpora in the catalog of the LDC, or of any corpus distribution agency for that matter, are dominated by a small number of linguistic genres and do not reflect the kind of up-to-date and balanced microcosm of language that the ANC is aiming to capture. Finally, in order to support multiple constituent research communities, the LDC is forced to be pluralistic about formats, whereas the ANC can provide a corpus and its annotations in a unified format, so that it is not necessary to customize every individual use of the data.

3.3.2 Data Acquisition and Distribution

All ANC data and annotations are distributed through the Linguistic Data Consortium (LDC). ANC annotations in stand-off documents that are linked to the primary data are also distributed via the ANC web site, since they are not covered by licensing restrictions.

Chris Cieri outlined the licensing situation for the ANC data, which are covered by two licenses pertaining to different portions of the ANC. Both licenses allow the data to be used freely for research purposes. The license for the Open Portion of the ANC (comprising about 80% of the data) does not contain any language restricting it to non-commercial use, nor does it prohibit re-distribution of the data. The license for the Restricted Portion (20%) restricts it to non-commercial use and allows no re-distribution.

Under current definitions of fair use, small segments of the texts (by convention something around 250 words) can be reproduced and re-distributed. The ANC licenses pertain to the ANC materials distributed by the LDC, but not to stand-off annotations created and distributed independently.

Much data that could be included in the ANC is available via the web and other sources under licenses such as the GNU public license and the Creative Commons "share-alike" license. So far, the ANC has avoided including these data in the corpus because the restriction that the data *must* be re-distributed under the same provisions of the original license is incompatible with both the ANC open and restricted licenses: in the former case, it is too restrictive, since the open license allows for completely unfettered redistribution; and in the second, it is too liberal by allowing re-distribution under any circumstances. Chris Cieri indicated that a third ANC license that mirrors the GPL and Creative Commons licenses could be added, thereby allowing the inclusion of a vast amount of available data from many different genres that had previously been excluded.

3.3.3 Output formats

The ANC should provide the corpus and annotations in formats compatible with widely used tools such as the Natural Language Tool Kit (NLTK) and UIMA. Making ANC available for inclusion in the NLTK will encourage future generations of students to use it. In addition, UIMA is gaining ground as standard; it has, for example, been adopted the DARPA GALE project as the platform for future annotation projects. Therefore, UIMA-compliant output would also encourage ANC use among members of the CL community.

3.3.4 Internationalization

The workshop participants emphasized that it must be guaranteed that the ANC is not "provincial" and can link to resource efforts in other countries. Annotation schemata need to be designed with an eye towards the international community.

The ANC uses the recommendations of ISO TC37 SC4's Linguistic Annotation Framework (LAF) to represent the corpus and its annotations. The LAF recommendations are being widely adopted in Europe and Asia, thus ensuring interoperability of the ANC with resources developed in these areas of the world.

Nancy Ide mentioned that in collaboration with the City University of Hong Kong, and with the involvement of representatives of WordNet, FrameNet, and European annotation

efforts, the ANC is co-organizing a conference to be held in November, 2007 in Hong Kong entitled "Towards Global Interoperability of Corpora and Resources". The goal of the conference is to bring together an international community of those involved in creating resources to learn about work across the globe and to discuss future directions to ensure compatibility, as well as to provide a survey of state-of-the-art techniques for new resource developers (especially for languages that have not yet developed significant resources of their own). The conference should serve as a means for the ANC to ensure interoperability with resources developed in other parts of the world.

It was also mentioned that multilingual annotations should be included, as they are helpful to the Machine Translation community. It was not clear what multilingual annotation of the ANC would involve, beyond the linking via WordNet sense tagging that would provide translation equivalents in other languages. One possibility would be to include materials in the ANC for which there exist annotated parallel translations (perhaps especially in languages relevant to North America such as French Canadian and Spanish), along the model of the MULTEXT-EAST Orwell parallel corpus.

3.3.5 Additional sources of data and annotations

Additional existing sources for annotated data were considered, such as the corpora used in the GALE project and the data being annotated for opinion at the University of Pittsburgh. The latter consists of translated FBI transcripts that are not necessarily representative of American English. The ANC is intended to include only "native" American English, to serve the needs of lexicographers developing American English reference works, corpus linguists studying the characteristics of American English, and developers of ESL textbooks and materials for teaching American English. (It was also noted that the Department of Education has indicated that it will use the ANC in the National Assessment of Adult Literacy for 2008, rather than using the BNC as it has done in the past.) It was suggested that the ANC could include English that is not necessarily produced by native speakers of American English[5], if these data were explicitly identified as such in the metadata description of the texts and users can select data on this basis. It was noted that ICSI/Berkeley has contributed a corpus of spoken transcripts to the ANC that identifies the native language of each speaker; where possible, a similar strategy could be applied to other data included in the ANC (or at least, the data could be identified as not reliably native American English).

The ANC has already set up a mechanism on its website through which anyone can contribute texts and/or annotations to the ANC. The ANC has so far received a number of works of fiction via this mechanism from various contributors. It has also received annotations for co-reference in a portion of the *Slate* magazine data, contributed by the University of Alberta, and annotation for part of speech using the CLAWS 5 and 7 tags (the tagset used for the BNC) from the University of Lancaster. The issue of the reliability of contributed annotations was raised, and it was suggested that the ANC publish a set of "best practice guidelines" for annotators. Use of such a set of guidelines

---

[5] The notion of a "native" speaker of American English is itself problematic; consultation with the American Dialect Society and the Linguistic Association of America led to no definitive definition. See http://americannationalcorpus.org/native-speaker.html.

by annotators intending to contribute to the ANC would also ease the transduction of contributed annotations into the ANC format.

Workshop participants identified several corpora (many already annotated) that could be contributed to the ANC:

- James Pustejovsky/Bran Boguraev: TimeBank
- Rebecca Hwa: Press releases from the White House
- Adam Meyers: blogs, wikipedia, political speeches
- Janyce Wiebe: FBI translations, annotated for subjectivity
- Oxford University Press: over two years's worth of journals and a corpus of New Yorker cartoon captions
- Becky Passonneau: spoken narrative data (Pear stories); same individual with the same story on 3 days (held by someone else)
- Jerry Hobbs/Kathy McKeown: multi-party meeting data

If the ANC becomes more visible within the US computational linguistics community, additional data may become available.

3.4 Annotation Strategies

3.4.1 Multiple annotations

Multiple annotations of ANC data are needed for continued computational linguistics research, including both annotations for multiple phenomena and multiple annotations of the same phenomenon from different theoretical perspectives.

Automatically-produced annotations, especially multiple annotations of the same type, are valuable for comparison and development of heuristics that can improve the performance of automatic annotation software. However, the determination of the workshop was that the most critical need is for a *manually-annotated sub-corpus* representing a range of genres and including annotations for a broad range of linguistic phenomena, as well as, in some cases, different annotations of the same phenomenon (e.g., FrameNet and PropBank semantic annotations). A manually-annotated sub-corpus would provide urgently needed training data to improve automatic annotation software, and, because it would be rendered in the ANC format, enable merging of different annotation types that can contribute to harmonization efforts such as the ULA.

It was generally agreed that a sub-corpus that has been manually annotated for WordNet senses and FrameNet frames is highly desirable. Annotation with WordNet senses and FrameNet frames has the further advantage that it will provide links to wordnets and framenets in other languages. For example, a WordNet sense-tagged lexical unit in the ANC will be automatically associated with its translations in the over 30 existing wordnets in other languages. Maximum value will be obtained if a portion of the sub-corpus consists of parts of the *WSJ* that include annotations for other phenomena (Penn Treebank syntactic annotation, Penn Discourse annotation, PropBank, NomBank, and TimeML).

3.4.2 Manually annotated sub-corpus

Ideally, manual annotations for multiple phenomena would be done for large, overlapping subset(s) of the ANC data across diverse genres, which could then be used to train and improve automatic annotation software in order to produce more reliable automatic annotations for the remainder of the ANC). This brings up several questions:

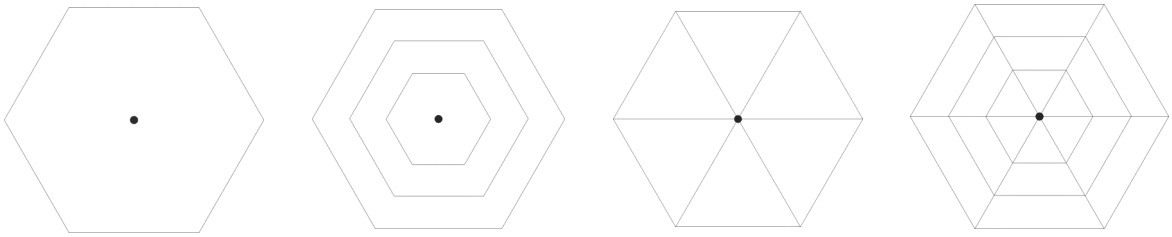
- How much data is enough for the sub-corpus? Different annotation types are required in different volumes for training. Should a common core be annotated for all phenomena?
- Which subparts or genres should be included in a manually-annotated sub-corpus? Different communities focus on different genres or domains—to what extent should the interests of a particular community govern the choice of genres to be included?
- How do we accommodate the need for example-based data sampling? Common annotation strategies identify difficult cases and annotate examples for the purposes of training annotation software. However, sufficient examples may not exist in the sub-corpus although they are present elsewhere in the ANC data.
- Which annotations should be included in a manually-annotated sub-corpus? Different communities are interested in different annotation types—to what extent should the interests of a particular community govern the choice of annotations? Some types of annotation can be performed automatically with high reliability; these annotation types can be validated rather than produced entirely by hand in the sub-corpus for relatively little cost. But other types of annotation are done well only when produced manually, and among these types, some require a far more intensive effort and are therefore more expensive and time-consuming to produce. In addition, different annotations are produced manually with varying reliability.
- Who will do the annotation? There are two possible scenarios: (1) manual annotations are created either by or for the ANC by qualified researchers, with significant quality control and cross-validation. The types of annotation to be done could then be chosen according to the needs of the scientific community. (2) Alternatively, the ANC can incorporate annotations that already exist or are created by others, which is more cost-effective but means that quality control may not be guaranteed, and the type of annotations that are done is opportunistic. The group discussed allowing a wide population of people to annotate over the web, but it was decided that untrained people would not be able to do the annotation reliably.

The workshop participants felt that the most cost-effective way for the ANC to proceed is to navigate between good science and practical constraints, by continuing to collect data from diverse genres in order to build a representative core, but at the same time to take advantage of previous and ongoing efforts by incorporating existing data and annotations that may or may not contribute to the genre balance of the core ANC. The most obvious example is the 1992 *Wall Street Journal* corpus, which has been and is being annotated for several linguistic phenomena. Because the *WSJ* represents a single domain, much of it would not be a part of the representative 100 million core of the ANC, but it can be included as part of the "varied" complement to the ANC core.

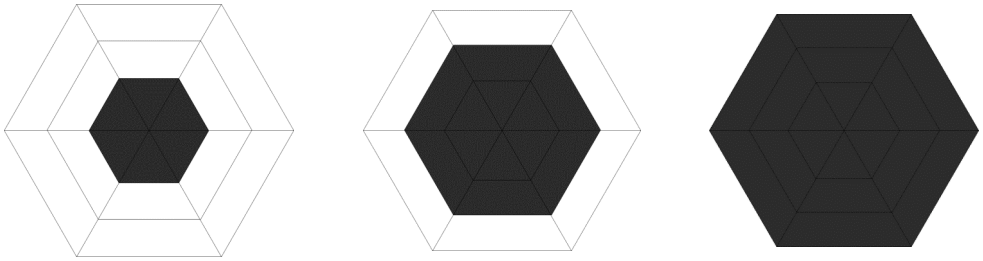
The "valued added" by incorporating such data into the ANC is that it would be rendered

into a common format; at present, annotations such as those in the *WSJ* exist in a variety of formats and require considerable effort to combine. To encourage additional annotations to be donated, the ANC (with the cooperation of the LDC) could offer the corpus to new annotation projects and students, and propose annotation projects to create overlaps and fill gaps.

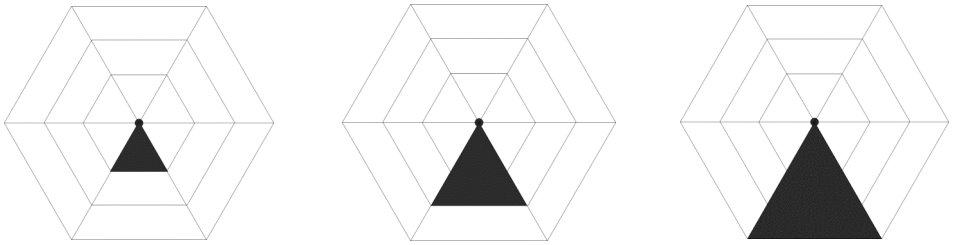
Chris Cieri offered the following analysis of sampling techniques that could be used to identify the manually annotated sub-corpus:



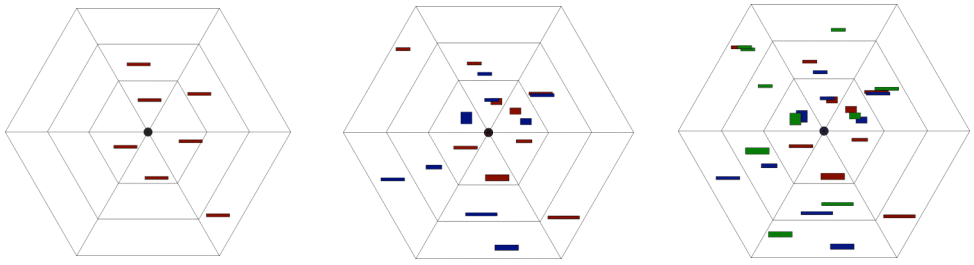
Core 100m word ANC Sampling by volume Sampling by genre/domain Sampling by both



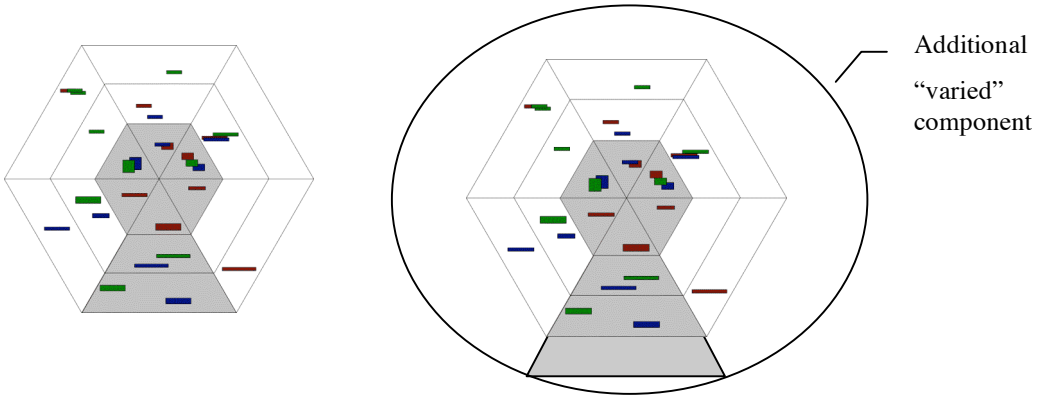
Increasingly large core samples



Increasingly large genre samples



Sampling by example (confidence)



Resulting core sample Resulting sample configuration including core + additional data

The Manually-Annotated Sub-Corpus could become a staple of CL research, to which researchers continually contribute additional annotations and data. As such, it could provide a baseline for annotation evaluation in the future.

3.4.3 Improving performance of automatic annotation software

We can approach the creation of a manually-annotated sub-corpus in several ways:

- manually annotate the sub-corpus from scratch, with no prior automatically-produced annotation as a basis
- manually correct the output of a single automatic annotation system in the sub-corpus
- apply several different automatic annotation systems for a given phenomenon, then manually correct annotations in the sub-corpus using the multiple results as a starting point

The manually-produced results will be used to (re-)train annotation software that can be applied to the full ANC, thus producing more accurate annotation for the whole corpus. Re-training can be optimized by applying strategies such as sample selection and active learning, wherein cases that are difficult for the learning algorithm are identified, and examples of these cases are then manually annotated and used to re-train the learning algorithm. If this strategy is applied at intervals throughout the manual annotation/correction process, the automatically-produced output should become increasingly reliable as the manual correction proceeds. Note that this type of strategy implies that (1) examples of difficult cases may not be numerous enough in the sub-corpus to provide adequate training data, and therefore examples from other portions of the ANC may be used (cf. "sampling by example" in the figures in the previous section); and (2) the sub-corpus and examples should cover multiple genres, so that examples are as learner/model independent as possible.

It is also possible to examine and compare results for the same phenomenon produced by different annotation systems to see if heuristics can be applied to automatically correct at least some portion of the annotations, prior to manual annotation. This strategy can also

be applied incrementally over the manual correction process, to produce increasingly accurate automatically-generated annotations.

3.4.4 Inter-annotator agreement

Inter-annotator agreement was considered to be an important consideration, especially for phenomena such as word senses where it is well known that agreement among humans is far from 100%. Inter-annotator agreement was also considered as valuable input for the active learning strategy described above, since cases in which the annotators disagree may help to identify cases that require more annotated examples to clarify.

The group felt that agreement values should be computed and provided with ANC annotations. Rebecca Passonneau suggested reporting multiple inter-annotator agreement statistics to present a more complete picture of agreement, because no one statistic is ideal for all datasets. She also suggested reporting other aspects of annotation results that would indicate which types of items were more difficult to achieve agreement on, and in cases of more than two annotators, which annotations were in greater agreement.

Given the creation of a subcorpus with multiple layers of annotation, and one that includes multiple genres, it was also suggested that the ANC include a new range of descriptive statistics that goes beyond the current frequency lists and bigram counts. This could begin with counts of annotation labels for the whole subcorpus and by genre, and counts of annotation label bigrams for each type of annotation. If merged annotations are rendered in the ANC graph format, descriptive statistics providing information about how the multiple layers interact could be produced by producing frequency counts for combinations of labels across layers.

3.5 Collaborations

The workshop participants agreed that increased collaboration among resource builders and annotators, as well as developers of annotation software, is an important step to avoid duplication of effort and ensure future interoperability. Recent efforts to devise standardized ways to represent language data and annotations in order to optimize its usefulness and flexibility (e.g., ISO TC37 SC4's Linguistic Annotation Framework (LAF), UIMA) and to harmonize annotation schemes for different phenomena on linguistic grounds (e.g., the Unified Linguistic Annotation effort and the corresponding workshops on Frontiers of Corpus Annotation) signal this as an opportune moment to establish collaborations among these different groups.

Another motivation for collaboration is that annotation practices, especially in terms of representation formats, have converged in recent years (e.g., the acceptance of stand-off annotation as the default standard). As a result, most recently developed formats and frameworks are virtually isomorphic; for example, the LAF representation developed by ISO TC37 SC4 is isomorphic to the format developed within UIMA, and rendering the ANC data (which uses LAF) into a format usable within UIMA is a relatively trivial exercise since both are based on feature structures. Collaboration will ensure that development in these and other projects continues to ensure interoperability among formats and tools.

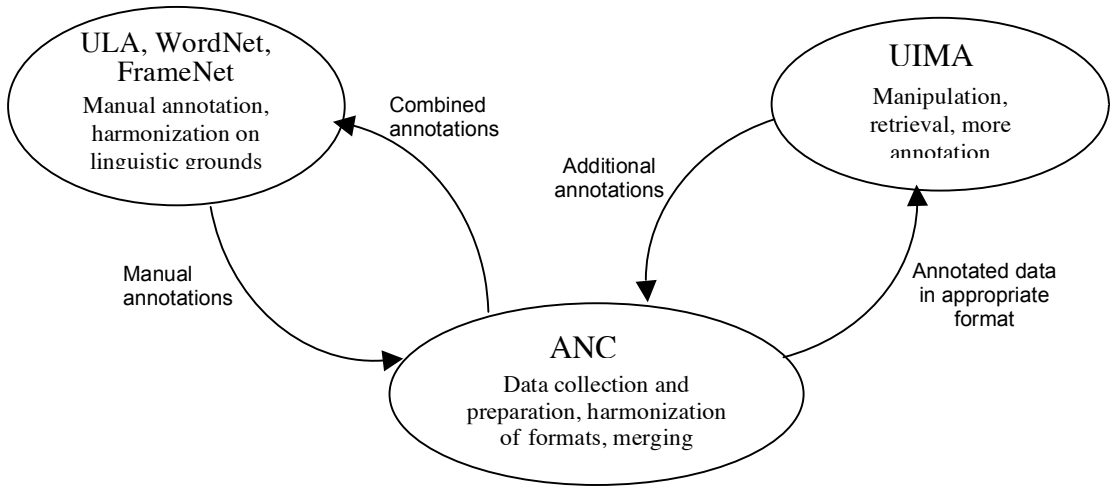
Specific collaborations were outlined as follows:

ANC-ULA: The ULA effort is manually annotating data for several phenomena, including PropBank (verbs-propositions), NomBank (nominalizations-arguments), TimeML (time, event structures), in order to examine interactions among different levels of annotation and, ultimately, devise a "unified linguistic annotation" scheme that harmonizes/enhances the linguistic information contained in each. This project needs data from different genres, preferably avoiding the costly effort to clean it up, and means to merge the different annotations so that interactions and overlaps can be studied. The ANC can provide clean, multi-genre data and has already developed means to merge annotations, and the results will be contributed for inclusion in the ANC.

The ULA project is also using data from Wikipedia, which, once annotated, may also be incorporated into the ANC. Potentially, ANC tools for data cleaning could be used to help the ULA project with data cleanup.

ANC-UIMA: A handler will be added to the ANC merge tool to generate ANC data and annotations in UIMA-compliant format. UIMA Common Analysis Structure (CAS) descriptions of annotations will be provided with all ANC annotations. Contributed annotations will be required to include a CAS.

IBM provides training and support for users and developers of UIMA. The ANC chief programmer will attend a UIMA training session, and additional collaboration, including potential ANC development of UIMA components, will be explored.



Collaborative inter-relationships

These collaborations bring together three groups with complementary expertise into a synergistic relationship that could lead to a major step in linguistic annotation.

Additional collaboration/cooperative efforts mentioned included: Center for Advanced Study of Language (University of Maryland), and GALE. Joseph Olive, representing the

GALE project at the workshop, was unsure that GALE data and annotations could be shared with the ANC.

3.6 Outreach

3.6.1 Resource clearinghouse

Workshop discussions led to the proposal that the ANC should serve as a clearinghouse for annotated data and provide a structured means to share information on relevant resources. The clearinghouse can be maintained on the ANC website. Another suggestion was to establish a wiki so that resources, annotated and non-annotated corpora, and software can be listed and made available if possible.

3.6.2 ANC Advisory Board

Because the ANC was initially funded by publishers, software vendors, and the NSF linguistics program, its Steering Committee and Advisory Board currently consist mainly of representatives from those communities. If the focus of ANC activity shifts to more directly serve the computational linguistics community, the ANC Steering Committee and Advisory Board should be re-formed to include members of this community. The participants proposed to form a new advisory board that will be actively engaged in ANC development, work towards reaching out to a broader set of communities, and better represent of the CL community in related areas, especially vis-a-vis representatives of professional organizations.

3.6.3 Workshops

The ANC has been involved in the organization of several workshops at major conferences over the past four years, including several recent editions of the NLPXML workshops at meetings of the EACL and ACL, and a workshop on Merging and Layering Linguistic Annotations held at LREC 2006. It was recommended that frequent workshops continue to be held in the future, possibly more directly focused on the ANC and its annotations and development, in order to ensure continued community input. In addition to holding workshops in conjunction with conferences such as ACL, EACL, and LREC, it was suggested that workshops or special sessions could also be held at meetings of the Linguistic Society of America and the Dictionary Society of North America. More direct ANC involvement in the series of Frontiers of Linguistic Annotation workshops was also suggested.

3.6.4 Additional activities

An ACL Interest Group (SIG) for annotation (SIGANN?) could be proposed.

Contact with related communities, such as the Semantic Web, the Knowledge Representation community, and ontologists should be made and maintained.

4 General Discussion on the Development of Linguistic Data and Annotations

The discussants stressed that annotation tasks are growing and becoming more complex; therefore, collaboration and coordination of all groups working on annotation is essential.

It was also stressed that annotation efforts should not duplicate, but build on existing resources and strategies. There has been considerable effort over the past 15-20 years in Europe to develop annotation standards, and this work can provide input and experience for future activities.

The fact that annotated resources need to be maintained was also emphasized. There is a history in both the US and Europe of resource development that is not followed up with funding to maintain and, where necessary, update the resource. This has led to a situation where resources have become obsolete, or, more often, become unavailable because developers have no support for distribution.

5 Next steps

The workshop participants agreed on the following immediate steps to be taken to achieve the identified goals:

(1) The ANC should propose to create a Manually Annotated Sub Corpus (MASC), which will provide:

   a. broad coverage of genres/domains, including newly emerging genres;

   b. validated annotations for WordNet senses and FrameNet frames, and, to the extent possible, existing manually-produced annotations for other phenomena (e.g. by including parts of the *Wall Street Journal*);

   c. measures of inter-annotator agreement, with confidence scores included with annotations;

   d. annotations in a layered, stand-off format, and mechanisms for merging all or selected annotations;

   e. the option to generate the corpus in a format conformant to UIMA and NLTK input specifications, and provide UIMA CAS descriptions for annotations.

(2) The ANC Advisory Committee should be re-formed to include representatives of the computational linguistics community;

(3) The ANC should add a third licensing category for ANC data through the LDC that allows for inclusion of data restricted by the "share-alike" provision, such as data licensed under the GNU Public license.

The participants also identified a set of longer-range activities that will contribute to the developments the community feels are necessary:

(1) The ANC should serve as a clearinghouse for annotated data;

(2) The ANC should provide a set of "best practice" guidelines for annotators, to guide future annotation projects and simplify the task of transducing annotations into the ANC stand-off format;

(3) The ANC should organize frequent workshops, both independent workshops and workshops held in conjunction with major conferences, including conferences of related groups (e.g., Linguistic Society of America, Dictionary Society of North

America)—the aim of the workshops would be both to inform and disseminate information and solicit community input.

6 Summary and Conclusions

The NSF workshop held in October, 2006 may prove to be a landmark event for linguistically-annotated corpora. The workshop brought together, for the first time, representatives of the major annotation and resource building projects in the U.S., together with researchers from both within and outside the U.S. who have been involved in international projects to develop and standardize language resources and evaluation techniques.

Perhaps the most impressive aspect of the meeting was the degree to and rapidity with which consensus among the participants was achieved. It was clear from the outset that the community is in full agreement that linguistically-annotated corpora are critical for the future development of language processing applications. In particular, all participants agreed that to advance the field, we need the following:

- **corpora representing diverse genres and topics,** rather than the domain-specific corpora that have been the focus of much previous activity—these corpora should also include **newly emerging genres** (e.g., newstalk, closed-caption text, blogs, wiki, etc.);
- **annotations for diverse linguistic phenomena,** in order to study their interactions and inter-dependencies;
- **annotations for the same phenomenon from different theoretical perspectives**, for the purposes of comparison and refinement;
- **a common representation format,** allowing annotations to be **layered** and allowing for **selective access** to annotations and the potential for **visualization**, in order to facilitate merging and comparison of annotations of diverse types as well as annotations for the same phenomena based on different theoretical approaches;
- **manually-annotated corpora** that can provide much-needed data for training and improving automatic annotation systems, together reliable information about linguistic phenomena upon which better language models can be built;
- **a clearinghouse and structured sharing of information** for relevant resources;
- **best practice guidelines** for annotators.

The group also made the following recommendations:

- **build upon existing annotated resources** where possible, rather than duplicating past work;
- **build resources with an eye toward the international community**, by designing/adapting schemata and formats to ensure interoperability with resources in other languages and/or developed outside the U.S.;
- **motivate annotation efforts in the context of NLP tasks**, rather than "annotate for annotation's sake";

- **provide confidence measures with annotations**, especially for annotations for which human agreement is problematic;
- **provide comprehensive meta-data with annotations**, including information concerning provenance, domain/genre, annotation categories, etc.

The consensus achieved by bringing together representatives of the different resource-building projects in the U.S. laid the foundation for increased collaboration among the various groups. Collaborative efforts have already begun as a direct or indirect result of the meeting: for example, the ULA group (PropBank. NomBank, TimeML) has adopted the ANC as one of the corpora it will collectively annotate; the ANC is working with the UIMA project to ensure interoperability and discuss potential moves toward a common framework; Rebecca Hwa has agreed to serve as a consultant to the ANC on machine learning and "active learning" techniques; the ANC Advisory Board is being re-formed and will meet at the ULA project workshop in Boston in February, 2007; representatives of the ANC and ULA projects are proposing the formation of an ANC SIG for annotation; and a joint NLPXML-FLAC workshop on annotation, encompassing the design and use of both linguistic annotations and annotation software and frameworks, will be held at ACL 2007 in Prague.

In addition to motivating collaborative efforts, the workshop participants identified several existing annotated corpora that can be contributed to the ANC. The ANC will (pending appropriate funding) transduce existing annotations into the ANC stand-off format, and enhance its tools for merging annotations to generate additional formats, including formats appropriate for UIMA and NLTK and a graph representation that can be used for both visualization and analysis.

The activities spawned by the NSF Workshop take a major step toward implementation of a long-term vision for the creation and use of linguistically-annotated corpora, involving both close collaboration and distribution of effort among the community. The ultimate aim would be to provide an expanding resource that is continually updated to answer community needs, which the community can collectively annotate over time. Through the ANC/LDC, all data and annotations will be freely distributed for research in a standard form, together with tools that enable merging and analysis without the repeated effort of re-formatting to serve individual needs. As best practices for annotation are identified and made available for community use, more and more annotations will be created in a consistent and fully usable way and will be compatible with resources and software developed outside the U.S. At the same time, the performance of automatic annotations systems will be enhanced, and systems to manipulate and exploit linguistic annotations will become both more powerful and easier to use. Such a scenario would provide the resources that are urgently required to support the next generation of major advances in natural language processing research.