An Open Linguistic Infrastructure for Annotated Corpora

Nancy Ide

1 Introduction

Annotated corpora are a fundamental resource for research and development in the field of natural language processing (NLP). Although unannotated corpora (for example, Gigaword, Wikipedia, etc.) are often used to build language models, annotations for linguistic phenomena provide a richer set of features and hence, potentially better models in the long run. It is widely accepted that a first step in the pursuit of NLP applications for any language is to develop a high quality annotated corpus with at least a basic set of annotations for phenomena such as part of speech and shallow syntax, while corpora for languages such as English, for which substantial annotated resources already exist, are increasingly being enhanced to include additional annotations for semantic and discourse phenomena (e.g., semantic roles, sense annotations, coreference, named entities, discourse structure). This is occurring for at least two reasons: first, more and deeper linguistic information, together with study of intra-level interactions, may lead to insights that can improve NLP applications; and second, in order to handle more subtle and difficult aspects of language understanding, there is a trend away from purely statistical approaches and (back) toward symbolic or rule-based approaches. Richly annotated corpora provide the raw materials for this kind of development. As a result, there is an increased demand for high quality linguistic annotations of corpora representing a wide range of phenomena, especially at the semantic level, to support machine learning and computational linguistics research in general. At the same time, there is a demand for annotated corpora representing a broad range of genres, due to the impact of domain on both syntactic and semantic characteristics. Finally, there is a keen awareness of the need for annotated corpora that are both easily accessible and available for use by anyone.

Nancy Ide
Vassar College, Poughkeepsie, New York USA e-mail: ide@cs.vassar.edu

Despite the need, there are very few richly annotated corpora, even for major languages such as English. This lack is most directly attributable to the high cost of producing such corpora. First, appropriate and, above all, available language data must be identified and acquired, often after lengthy copyright negotiations or painstaking web search for data unfettered by licensing limitations. Preparation of the data for annotation is notoriously difficult, especially when data come in a variety of formats, each of which must be cleaned to remove formatting information or, in the case of web data, extensive amounts of interspersed HTML (even more difficult if the format needs to be preserved); differences in character sets also have to be resolved in this step. Once prepared, annotation software may be applied to provide a base for manual validation, or annotations may be performed manually from the start; in either case, some environment for accomplishing the manual work must be provided. To be maximally useful, manual validation or annotation must be performed by multiple annotators and under controlled circumstances. For annotations at the semantic or discourse level, such as sense tagging or coreference, considerable effort to ensure the quality of the manual work must be expended, for example, by computing inter-annotator agreement metrics. Thus, corpus development can require several man-years of labor-intensive effort and, correspondingly, substantial funding. But while there has been some support for corpus creation and development over the past two decades, especially in Europe, in general the substantial funding required to produce high quality, richly annotated corpora, can be relatively difficult to acquire. Furthermore, the production and annotation of corpora, even when they involve significant scientific research, often do not, *per se*, lead to publishable research results. It is therefore understandable that many researchers are unwilling to get involved in such a massive undertaking for relatively little reward.

One means to offset the high cost of corpus creation is to distribute effort among members of the research community, and thereby distribute the cost as well. To this end, the American National Corpus (ANC) project[1] undertook to provide data and linguistic annotations to serve as the base for a collaborative, community-wide resource development effort (the ANC Open Linguistic Infrastructure, ANC-OLI) [12]. The fundamental premises of the effort are, first, that all data and annotations must be freely available to all members of the community, without restriction on use or redistribution, and second, that once a base of data and annotation was established, the resources would grow as community members contributed their enhancements and derived data. To ensure maximum flexibility and usability, the project has also developed an infrastructure for representing linguistically annotated resources intended to solve some of the usability problems for annotations produced at different sites by harmonizing their representation formats. We describe here the resources and infrastructure developed to support this collaborative community development and the efforts to ensure full community engagement.

[1] www.anc.org

2 Requirements for a Collaborative Annotation Effort

To be successful, an effort to involve the language processing community in collaborative resource development must meet several requirements in order that the resources meet community needs, and contribution of data and annotations as well as use of the available resources is easy for community members. Building on discussions held at a U.S. National Science Foundation (NSF)-sponsored workshop held in Fall, 2006[2], we identified the following general criteria for a collaborative community annotation effort for the field.

2.1 Open Data

In order to ensure that the entire community, including large teams as well as individual researchers, has access and means to use the resources in their work, all data and annotations included in the ANC-OLI should be either in the public domain or under a license that does not restrict redistribution of the data or its use for any purpose, including commercial use. (e.g., the Creative Commons Attribution (CC-BY) license[3]. Data under licenses such as GNU General Public License[4] or Creative Commons Attribution-ShareAlike[5] should be avoided because of the potential obstacle to commercial use imposed by the requirement to redistribute under the same terms.

2.2 Data Diversity

The lack of diverse data to support NLP research and development is well-known within the community. Even today, the corpora most frequently used by the community are the Penn Treebank corpus, the Chinese Treebank, EuroParl, and Wikipedia.[6], all of which are either very skewed for genre and/or unannotated. This is a result, of course, of the labor required to obtain large amounts of broad genre open data that can be annotated *and* redistributed with its annotations. The ANC-OLI should therefore include data from a range of different written and spoken genres, includ-

[2] The NSF workshop, held October 29-30, 2006, included the following participants: Collin Baker, Hans Boas, Branimir Bogureav, Nicoletta Calzolari, Christopher Cieri, Christiane Fellbaum, Charles Fillmore, Sanda Harabagiu, Rebecca Hwa, Nancy Ide, Judith Klavans, Adam Meyers, Martha Palmer, Rebecca Passonneau, James Pustejovsky, Janyce Wiebe, and funding organization representatives Tatiana Korelsky (NSF) and Joseph Olive (DARPA). A report summarizing the consensus of the workshop participants is available at http://anc.org/nsf-workshop-2006.

[3] creativecommons.org/licenses/by/3.0/

[4] www.gnu.org/licenses/gpl.html

[5] creativecommons.org/licenses/by-sa/2.5/

[6] Based on entries in the LRE Map, http://www.resourcebook.eu/LreMap/faces/views/resourceMap.xhtml

ing but not limited to the genres in "representative" corpora such as the Brown Corpus and the British National Corpus. It should also include topic-specific data and newer genres unrepresented in older language data collections, such as tweets, blogs, wikis, email, etc. Although modalities other than text should be the focus at the start, in principle the ANC-OLI should include and support audio, image, and video as well.

2.3 Annotation Types

The ANC-OLI should include automatically-produced annotations, especially annotations of the same phenomenon, which are valuable for comparison and development of heuristics that can improve the performance of automatic annotation software. In addition, there is a critical need for data that is manually-annotated for a broad range of linguistic phenomena, in order to provide much needed training data to improve automatic annotation software and machine learning. The ANC-OLI should seek support for manual validation of a (possibly small) sub-component of its holdings, but we expect to rely heavily on community contributions to provide high quality, manual annotations. In general, the production of annotations should be application-driven (e.g., discourse level annotations useful to Question Answering).

Whether automatically or manually produced, annotations should represent 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. Manual annotations over the same data utilizing widely used lexical and semantic resources such as WordNet senses and FrameNet frames are valuable as a step toward harmonizing such resources, which is a critical need for the field. Use of WordNet and FrameNet has the further advantage that it provides links to wordnets and framenets in other languages; for example, a WordNet sense-tagged lexical unit is automatically associated with its translations in the over thirty existing wordnets in other languages.

Ensuring the compatibility of annotation semantics–i.e., the linguistic categories used to describe the data–is still an area for research, and no attempt should be made by the ANC-OLI to resolve it. Rather, the ANC-OLI should encourage and contribute to efforts to devise means to harmonize linguistic annotation categories such as the ISO TC37/SC4 Data Category Registry (ISOCat [18]), General Ontology of Linguistic Description (GOLD, [7] and Ontologies of Linguistic Annotation (OLiA [3]), and in general foster the movement toward "semantics by reference" wherein the definition of a linguistic category used in an annotation is provided by referencing the URI of the category in question.

2.4 Format

To be successful, an effort to involve the community in a collaborative resource development effort must ensure that it is easy for community members to contribute and that the resulting resources are easy for community members to use, both individually and together. In the past, widely used corpora have been enhanced by community members, and in some cases the added resources have been made publicly available, but the lack of consistency among formats has prevented combined use of the existing and added annotations. The most obvious case in point is the one million word *Wall Street Journal* corpus known as the Penn Treebank [19], which over the years has been fully or partially annotated for several phenomena over and above the original part-of-speech tagging and phrase structure annotation. The usability of these annotations is limited, however, by the fact that most of them were produced by independent projects using their own tools and formats, making it difficult to combine them in order to study their inter-relations.

The obvious obstacle to the combined use of annotations produced at different sites is the lack of standards for representing linguistically annotated language data, including not only annotated corpora but also lexicons, treebanks, propbanks, etc. The ANC-OLI should address this obstacle as broadly as possible, by seeking a solution that would cover the greatest number of situations in the short term, and at the same time serve over the long term as a viable approach to the multiple formats problem.

Transducing among different formats, especially complex formats such as the Penn Treebank syntax and PropBank semantic role annotations that depend on it, is often non-trivial. Therefore, ANC-OLI contributors cannot be expected to expend resources to provide their annotations in any format other than the one their in-house tools produce, and users cannot be expected to adapt ANC-OLI annotations for use with either in-house or off-the-shelf tools. However, to be usable together, both internally-produced and contributed annotations must be represented in a single, usable format. This format must be both powerful and generic enough to allow annotations in any representation (e.g. LISP structures, XML) and with any internal structure (e.g., tree, graph) to be readily *mapped* to it without information loss, and flexible and standardized enough to enable linking to resource efforts in other areas of the world. For ease of use, ANC-OLI data and annotations should also be made available not only in a common format, but also in formats compatible with widely used tools such as the Natural Language Tool Kit (NLTK), GATE, and UIMA, as well as other commonly used formats such as the Resource Description Format (RDF) and the IOB format used in CoNLL shared tasks.

2.5 Access

Access should be easy and open via the web. Selective access should also be provided, so that users can choose to download only the annotations and data of interest

to them, in a format that is convenient for their purposes. In addition, there should be tool support for the data and annotations in the common format.

2.6 Maintenance

There must be provision for maintenance and sustainability. 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. Therefore, to ensure sustainability of the resource, the resources should be made available through a major data center such as the LDC, which can guarantee long term availability of the resource. In the short term, availability through a major data center will increase the visibility and accessibility of the resources.

2.7 Coverage

The ANC-OLI is based on the American National Corpus, which by definition contains only American English data. The ANC-OLI should be expanded to include other languages and media such as audio, video, image, etc. at the earliest possible time.

2.8 Fostering community involvement

The idea of an annotated resource deliberately intended for collaborative development is a relatively new one in the field. Until recently, the addition of annotations to common data by different individuals or groups was done in an *ad hoc*, uncoordinated way, and there was never a clear intention to use the annotations together or even share them with the rest of the community. The growing promotion of sharable, "open" resources over the past few years (largely engendered by the open software movement) has created a major shift in community perspective concerning the need to accommodate more universal usability of resources and tools, but in general, the *de facto* scenario in people's minds does not include giving resources to another individual or group for their use. Therefore, there is, as yet, no collective mentality fostering collaborative resource development, although this is clearly on the horizon. In the meantime, to engage the community and perhaps move them more quickly toward adoption of the collaborative model, it is necessary to familiarize researchers and developers with the premises behind collaborative development and promote its adoption.

3 ANC-OLI

3.1 History

The American National Corpus project was launched in 1998 [9], motivated by developers of major linguistic resources such as FrameNet[7] and Nomlex[8], who found that usage examples extracted from the 100 million word British National Corpus (BNC), the largest corpus of English across several genres available at the time, were often unusable or misrepresentative for developing templates for the description of semantic arguments and the like, due to significant syntactic differences between British and American English. The ANC project was originally conceived as a near-identical twin to its British cousin: the ANC would include the same amount of data (100 million words), balanced over the same range of genres and including 10% spoken transcripts just like the BNC.

The BNC was substantially funded by the British government, together with a group of publishers who provided both financial support and contributed a majority of the data that would appear in the corpus. Based on this model, the ANC looked to similar sources, but gained the support of only a very few U.S. publishers and a handful of major software developers, who provided about $400,000 to support the first four years of ANC development, an order of magnitude less funding than that which supported development of the BNC.

British publishers provided the bulk of the data in the 100 million word BNC. The plan for the ANC was that the sponsoring publishers and software vendors would do the same for the ANC. However, only a very few of the ANC supporters eventually contributed data to the corpus.[9] As a result, it was necessary to attempt to find data from other sources, including existing corpora such as the Indiana Center for Intercultural Communication (ICIC) Corpus of Philanthropic Fundraising Discourse, and the Charlotte Narrative and Conversation Collection (CNCC), together with government documents, biomedical articles, and other public domain material on the web.

In 2003, the ANC produced its first release of eleven million words of data, which included a wide range of genres of both spoken and written data . Annotations included word and sentence boundaries and part-of-speech annotation produced by two different taggers in standoff form, that is, provided as separate files with links into the data.[10] To our knowledge, the ANC First Release was the first large, publicly available corpus to be published with standoff annotations. In 2005, the ANC released an additional eleven million words, bringing the size of the corpus to twenty-two million words. The Second Release includes data from additional genres, most notably a sizable sub-corpus of blog data, biomedical and technical

[7] www.icsi.berkeley.edu/ framenet

[8] nlp.cs.nyu.edu/nomlex/index.html

[9] The consortium members who contributed texts to the ANC are Oxford University Press, Cambridge University Press, Langenscheidt Publishers, and the Microsoft Corporation.

[10] The contents of the ANC First Release are described at http://www.anc.org/FirstRelease/

reports, and the *9/11 Report* prepared by the U.S. Government. The Second Release was issued with standoff annotations for the same phenomena as in the First Release, as well as annotations for shallow parse (noun chunks and verb chunks). Notably, the ANC Second release also included the first community contributed annotations of the corpus: manually produced coreference annotation of about 100,000 words of Slate magazine articles contributed by University of Alberta, and two additional part of speech annotations using the CLAWS 5 and 7 tags used in the BNC contributed by University of Lancaster.

In 2006, the project made fifteen million of the ANCs twenty-two million words that were not restricted for any use available for download as the "Open ANC (OANC) from the ANC website.[11] The fully open distribution model pioneered by the OANC has now been adopted for all future releases of data and annotations[12] It was at this point that the ANC-OLI was conceived [12], thus creating the first collaborative, community-wide resource development effort in the field. Since then, three syntactic parses of eleven million words of the OANC (using the Charniak and Johnson parser, MaltParser, and LHT dependency converter, respectively) and named entity annotations of the entire OANC produced by the BBN tagger [20], have been contributed.

The next year, the ANC project received a substantial grant from the U.S. National Science Foundation[13] to produce a half-million word Manually Annotated Sub-Corpus (MASC) of the OANC that would include automatically-produced annotations for logical structure (paragraph, section, headings, etc.), word and sentence boundaries, part of speech and lemma, shallow parse, and named entities, and to manually add annotations for WordNet senses and FrameNet frames to portions of the corpus. From the outset, the project was designed to serve as a centerpiece for the ANC-OLI, and so to facilitate initial community contribution, materials for the MASC were drawn from sources that have already been heavily annotated by others (where licensing permitted). MASC currently includes a 50K subset consisting of OANC data that has been previously annotated for Penn Treebank syntax, PropBank predicate argument structures, Pittsburgh Opinion annotation (opinions, evaluations, sentiments, etc.), TimeML time and events and several other linguistic phenomena. It also includes a handful of small texts from the so-called Language Understanding (LU) Corpus[14] that was annotated by multiple groups for a wide variety of phenomena, including events and committed belief; and 5.5K words of Wall Street Journal texts that have been annotated by several projects, including Penn Treebank, PropBank, Penn Discourse Treebank, TimeML, and the Pittsburgh Opinion project. All of these annotations, apart from 420K of annotations for Penn Treebank syntax[15], were contributed to the project.

---

[11] www.anc.org/OANC/index.html

[12] However, since 2005 the ANC project had no funding for production of additional data.

[13] NSF CRI 0708952

[14] MASC contains about 4K words of the 10K LU corpus, eliminating non-English and translated LU texts as well as texts that are not free of usage and redistribution restrictions.

[15] The MASC project commissioned the remainder of the annotation from the Penn Treebank project.

The first full version of the corpus was released in 2012, including a separate sentence corpus [23] that provides sense-tags for approximately 1000 occurrences of each of 114 words chosen by the WordNet and FrameNet teams (ca. 114,000 annotated occurrences).

3.2 Meeting the Requirements for Community Collaboration

3.2.1 Open Data and Data Diversity

The requirement for *open data* imposes severe limits on what can be included in the corpora distributed by the ANC-OLI, making data acquisition the major issue for ANC-OLI development. Over the past five years we have gathered approximately 50 million words of open data[16], not including public domain data that can be acquired from government sites and web archives of technical documents. While these latter sources can provide virtually limitless amounts of data, the requirement for *data diversity* means that acquisition efforts must focus on other data types, especially those that are rarely published as open data such as fiction, tweets, etc. The OANC contains about 3 million words of spoken data spoken data (face to face, telephone conversations, academic discourse), and over 11 million words of written texts (government documents, technical articles, travel guides, fiction, letters, non-fiction). The contents of the MASC corpus are given in Table 1.

To date, the ANC-OLI has gathered open data from the following sources:

1. Contributions from publishers who are willing to provide data under a non-restrictive license, including non-fiction materials donated to the ANC by Oxford University Press and Cambridge University Press, travel guides from Langenscheidt, and SLATE magazine articles from Microsoft. To protect their interests, publishers sometimes provide only a subset of a complete book or collection.
2. Web materials in the public domain or licensed under non-viral licenses such as CC-BY. Government documents and debate and court transcripts, as well as technical articles in collections such as Biomed Central[17] and the Public Library of Science[18], are typically in the public domain. Although more difficult to track down, blogs, fiction, and other writing such as essays are very often distributed over the web under licenses such as CC-BY.
3. Contributions from college students of class essays and other writing. College students produce considerable volumes of prose during their academic careers, and very often this data is discarded or forgotten once handed in to satisfy an assignment. The ANC-OLI provides a web interface for contributions of this kind that includes a grant of permission to use the contributed materials[19].

[16] Lack of funding for processing the data currently prevents its publication.

[17] www.biomedcentral.com

[18] www.plos.org

[19] www.anc.org/contribute.html

Genre	No. files	No. words	Pct corpus
Court transcript	2	30052	6%
Debate transcript	2	32325	6%
Email	78	27642	6%
Essay	7	25590	5%
Fiction	5	31518	6%
Gov't documents	5	24578	5%
Journal	10	25635	5%
Letters	40	23325	5%
Newspaper	41	23545	5%
Non-fiction	4	25182	5%
Spoken	11	25783	5%
Technical	8	27895	6%
Travel guides	7	26708	5%
Twitter	2	24180	5%
Blog	21	28199	6%
Ficlets	5	26299	5%
Movie script	2	28240	6%
Spam	110	23490	5%
Jokes	16	26582	5%
TOTAL	376	506768	

Table 1. Genre distribution in MASC

4. Direct solicitation for use of web materials. We have on occasion identified a web site containing interesting or substantial materials and contacted the relevant parties directly to explain our use of the data and ask for permission to use it. We have also contacted providers whose data are freely available for access to the materials in a form more manageable for processing purposes. So far, none of our requests has been turned down.
5. Contributions from colleagues in the field and data centers such as the Linguistic Data Consortium (LDC)[20]. We have received data contributions, including significant amounts of spoken data, from several NLP and linguistics projects, including the Indiana Center for Intercultural Communication (ICIC) Corpus of Philanthropic Fundraising Discourse[21], Project MUSE's Charlotte Narrative and Conversation Collection (CNCC)[22], the Michigan Corpus of Academic Spoken English (MICASE)[23], and the International Computer Science Institute (ICSI) Meeting Corpus [16]. We have also received contributed annotations from the

[20] www.ldc.upenn.edu

[21] liberalarts.iupui.edu/icic/research/corpus_of_philanthropic_fundraising_discourse

[22] newsouthvoices.uncc.edu/

[23] http://quod.lib.umich.edu/m/micase/

Penn Treebank project, the PropBank project, the Pittsburg Opinion annotation project, TimeBank, and several others.

Acquisition of almost all of these data was non-trivial, requiring substantial time and effort to solicit contributions from publishers, projects, and even college students, and to identify suitably open materials on the web. Contributions from the research community at large have also so far been relatively meagre, typically due to licensing constraints. As awareness of the nature of and need for open data increases, these contributions are more and more readily forthcoming.

3.2.2 Annotations

The 15 million word OANC includes automatically-produced annotations for logical structure, sentence and token boundaries, part of speech and lemma (4 different taggers and tag sets), noun chunks, verb chunks, and named entities. Eleven million words are automatically annotated for two dependency parses and one phrase structure parse. MASC contains a richer set of annotations, all manually produced or hand validated, over all or parts of the corpus as shown in Table 2. The MASC Sentence Corpus consists of approximately 110,000 sentences with WordNet sense annotations for 114 words. The sentences include every occurrence of each of the 114 words in MASC together with occurrences drawn from the OANC to fill out the balance of 1000 sentences per word.

While every effort has been made to include as diverse a set of annotations, including multiple annotations of the same type representing different theoretical approaches, the ANC-OLI does not have the resources to produce the full range of possible types, especially for the MASC data which requires manual validation. One particular lack is a dependency parse of MASC, which would provide a complement to the Penn Treebank phrase structure analysis. Discourse-level annotation of a variety of types would also be desirable for MASC. We will rely on community collaboration and contribution to fill these gaps.

Automatic annotation of OANC data is easier to produce, but still requires programming effort to render into GrAF. Also, the accuracy of automatically produced annotations over OANC data tends to degrade severely, since most annotation software is trained on a single or relatively constrained set of genres, whereas the OANC data is far more varied. Hopefully, the availability of diverse data will spark experimentation with the impact of domain and genre on the performance of automatic annotation software.

3.2.3 Format

The representation format of the ANC-OLI annotations must serve two purposes: it must be possible to transduce from formats of contributed annotations to the ANC-OLI format without loss of information, and the format must be interoperable with diverse tools and frameworks for searching, processing, and enhancing

Annotation type	No. words
Logical	506659
Token	506659
Sentence	506659
POS/lemma (GATE)	506659
POS (Penn)	506659
Noun chunks	506659
Verb chunks	506659
Named Entities	506659
FrameNet	39160
Penn Treebank Syntax	506659
PropBank	55599
Opinion	51243
TimeBank	55599
Committed Belief	4614
Event	4614
Dependency treebank	5434
Coreference	506659
Discourse segments	506659

Table 2: Summary of MASC annotations

the corpus. For this reason, the representation of all ANC-OLI annotations follows the specifications of the International Standards Organization (ISO) Linguistic Annotation Framework (LAF) [15], which provides a framework for representing annotations based on an abstract model consisting of a graph of features structures, two very powerful and general data structures that have been widely used, either directly or as an underlying model, to represent linguistic information [11]. A fundamental tenet of the LAF model is that all annotations are in stand-off format, with references to primary data or other annotations.[24]. The Graph Annotation Format (GrAF) [13, 14], the XML serialization of the model, is intended to function in much the same way as an interlingua in machine translation, that is, as a "pivot representation into and out of which user- and tool-specific formats are transduced, so that a transduction of any specific format into and out of GrAF accomplishes the transduction between it and any number of other GrAF-conformant formats. The rendering of ANC-OLI data and annotations in GrAF thus satisfies the criteria outlined above: it is powerful enough to represent annotations contributed in any format, easy to transduce ANC-OLI annotations to other formats, and it conforms to a widely adopted international standard. The graph-based format also enables trivial merging of annotations rendered in GrAF. Furthermore, the generality of the abstract model makes mapping to formats such as the Resource Description Format

[24] Allowing annotations to reference other annotations differentiates GrAF from other representation formats, such as Annotation Graphs [2]

(RDF) [17], which is the format used in the Semantic Web, and the UIMA Common Analysis System (CAS) [8].

The generic graph model underlying GrAF is isomorphic to that of emerging Semantic Web standards, notably RDF/OWL, thus making conversion between GrAF and RDF/OWL representations trivial. The GrAF representation of MASC has recently been rendered into POWLA [4], an RDF/OWL linearization of PAULA, a generic data model for the representation of annotated corpora [6, 5]. This representation includes linkage of its WordNet and FrameNet annotations to RDF instantiations of these resources, as well as linkage of linguistic categories used in MASC's other annotation layers to types in the POWLA OWL/DL ontology.[25] The RDF/OWL instantiation opens up the potential to formulate queries that combine information from the linked versions of WordNet, FrameNet, and MASC using an RDF query language such as SPARQL [25]. The RDF/OWL version of MASC is publicly available as a part of the Linguistic Linked Open Data cloud[26].

The ANC-OLI project is committed to rendering contributed annotations into GrAF. To date, all of the annotation types in Table 3.2.2, which came to us in a variety of both stand-off and embedded (in-line) formats, have been rendered into GrAF without information loss. The transduction process is not always trivial; for example, to be transduced to GrAF standoff form, in-line annotations must first be extracted from the text and then realigned to refer to the primary data document. Another problem results from variations in tokenization among the different annotations; this is solved by GrAF's provision for a segmentation document that defines minimally granular regions over a primary resource, which may then be combined (if necessary) and referenced by different tokenizations. The difficulties encountered in the transduction process typically arise from inconsistencies or omissions in the original format, which must be rectified in the GrAF representation. The problems are at any rate informative for the development of best practice annotation guidelines.

3.2.4 Access and maintenance

All ANC-OLI data are freely downloadable from the web, without the need to sign a license or provide any information. In addition, to ensure sustainability of the resources, all data and annotations are held and distributed by the Linguistic Data Consortium for no cost.

For use of the available resources, the ANC-OLI has developed an open source GrAF API[27] for reading and writing GrAF files and also provides a web application, called ANC2Go, that enables a user to choose any portion or all of MASC and the OANC together with any of their annotations to create a "customized corpus". The customized corpus can be delivered in any of several formats, including inline XML

[25] For more details, see Chiarcos, et al., in this volume.

[26] linguistics.okfn.org/llod

[27] http://sourceforge.net/projects/iso-graf/

(input to any XML-aware program, including BNC's XIARA which then allows for comparative studies), token/pos (input to commonly used concordancing software), CONLL IOB format, tagged input for the Natural Language Toolkit (NLTK), and RDF. The project also provides modules to import/export from the widely used annotation and analysis frameworks GATE[28] and UIMA, so that ANC-OLI annotations are directly usable in these systems. However, in addition to being readily transduced to other formats, the GrAF format is useful in itself: one of the most salient features of the graph representation for linguistic annotations is the ability to exploit the wealth of graph-analytic algorithms for information extraction and analysis. For example, it is trivial to merge independently-produced annotations of the same data in GrAF form, as well as to apply algorithms to find common sub-graphs that reflect relations among different annotations.

3.2.5 Coverage

Because the ANC-OLI grew out of the American National Corpus project, the included corpus resources currently include only American English spoken transcripts and written texts. Ideally, the project should expand to cover other modalities, including speech (audio), video, and image, as well as other languages. To address the lack of multilingual data in the ANC-OLI, we have recently launched MultiMASC [10], which builds upon MASC by extending it to include comparable corpora in other languages. Here, comparable means not only representing the same genres and styles, but also including similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data and expand the collaboration effort upon which it depends. The eventual result is envisaged to be a massive, multi-lingual, multi-genre corpus with comparable multi-layered annotations that are inter-linked via reference to the original MASC or, perhaps more interestingly, to the RDF/OWL instantiation of MASC and associated resources described in Chiarcos, et al. (this volume).

3.2.6 Fostering community involvement

Following the familiar quote "build it and they will come"[29], by virtue of their existence and availability, community use of the OANC and MASC has been immediate and substantial. Contribution of annotations, on the other hand, has been slower to develop but is now beginning to gain momentum. In the first years of OANC availability (2005 onward), only a handful of annotations were contributed, including the output of three different parsers[30] and named entity annotation produced by the BBN Tagger [21]. MASC has enjoyed better success, in large part because it is both

[28] General Architecture for Text Engineering; http://gate.ac.uk

[29] Taken from *Field of Dreams*; see http://en.wikipedia.org/wiki/Field_of_Dreams

[30] The Charniak and Johnson (2005) parser, MaltParser, and LHT dependency converter.

a newer resource and one that has been more widely publicized within the community via conference papers and workshops. The first release of 82K includes a 50K subcomponent for which several annotation layers were contributed, including Penn Treebank syntax, PropBank semantic roles, TimeML time and event annotation, and Pittsburgh opinion annotations. Additional annotations of MASC data for spatial information, PropBank semantic roles, discourse (Penn Discourse Treebank), and "deep semantics" (Groningen Meaning Bank), among others, are underway. We also expect that MASC–either the corpus and some or all annotations, or the sense-tagged sentence corpus–will be used in upcoming SemEval exercises[31].

Collaborative community development goes beyond the contribution of annotations. Such development crucially relies on the community to identify errors in order to continually improve the resources, together with contribution of derived data such as frequency lists, ngrams, statistics reflecting the distribution of various phenomena, etc. Another important development activity involves the incorporation of ANC-OLI data and annotations into platforms and frameworks that enable others to work with them, beyond those already provided by the ANC project itself. Currently, community members have spontaneously taken up incorporation of MASC into the OpenNLP machine learning toolkit[32] and development of a corpus reader for ANC-OLI data and annotations for the Natural Language Toolkit (NLTK), two important frameworks for NLP research and education.

As noted earlier, collaborative development is not yet in the mainstream of activity within the language processing community, and so it is still necessary to promote community involvement through publicity at conferences and workshops, together with the use of OANC and, in particular, MASC, in shared tasks such as CONLL, SemEval, and *SEM. It will require a significant shift in the community mindset before its members reflexively contribute their annotations of ANC-OLI data and derived information, given the established practice in the field of *consuming* resources with no expectation of return, a practice most evident in the procedures of resource repositories such as LDC and ELRA.[33] Widespread acceptance of the collaborative resource development model is exacerbated by the fact that preparing annotations and derived data for use by others can require additional and sometimes considerable effort. Nonetheless, recognition that the need for richly annotated and inter-linked language resources can be most efficiently met through a collaborative community development effort is increasingly widespread and motivates numerous national and international funding programs aimed at infrastructure development for NLP research.

---

[31] http://aclweb.org/aclwiki/index.php?title=SemEval_Portal

[32] http://opennlp.apache.org

[33] Such repositories were set up to answer the call for resource reusability which, no doubt in large part because information added to these resources was until recently unlikely to be usable by others, always referred to the consumer-only model.

4 ANC-OLI in Context

The ANC-OLI corpora provide a unique resource in terms of both their content and configuration, as well as the collaborative aspect of their development. For example, the two standard broad genre corpora for English, the Brown Corpus and the British National Corpus (BNC), provide only part of speech annotations, in contrast to the richer set of annotations in the OANC and particularly in MASC. In addition, Brown and BNC include only data produced prior to widespread use of the web, which has radically affected lexical and syntactic usage and fostered the emergence of new genres. The one million word *Wall Street Journal* corpus known as the Penn Treebank [19] has been fully or partially annotated for several phenomena beyond the original part-of-speech tagging and phrase structure annotation over the years, but most were produced by independent projects using their own tools and formats, making it difficult to use these annotations together. Of course, the lack of genre diversity of this corpus, which contains texts from a single domain that have been edited to conform to a consistent "Wall Street Journal style", is well known as a major drawback for its use in training language models for broad-range syntactic and semantic phenomena.

The corpus closest to ANC-OLI in terms of richness of annotation and currency of language is the one million word English OntoNotes corpus [24], which includes annotations for Penn Treebank syntax, sense annotations using an in-house sense inventory, PropBank predicate argument structures, coreference, and named entities represented in a "normal form". As in MASC, all annotations have been hand-validated. However, the OntoNotes corpus represents a limited set of genres (newswire, broadcast news, and broadcast conversation), and, because of the need to compile annotations into the internal OntoNotes database, annotations produced by others cannot be added to the corpus. Also, unlike ANC-OLI data, OntoNotes is restricted for research use only and requires licensing through the LDC.

Very recently, two collaborative annotation efforts have been initiated that share some aspects of ANC-OLI development. One, the *Language Library*[34], asks community members to apply their software to provided data and contribute the results. The Language Library data are for the most part freely available, although inclusion of large amounts of multi-lingual Wikipedia data imposes the "share-alike" restriction that typically prevents its use for commercial purposes. The inspiration for the Language Library came directly from the ANC-OLI collaborative model, with the intent to expand the coverage to multiple languages. The new MultiMASC effort will extend the ANC-OLI to other languages, but will differ from the Language Library because the data will represent a broad range of genres, include only manually produced or validated annotations, ensure all annotations are represented in a harmonized format, and, by virtue of the common format, enable inter-linkage of linguistic phenomena at all levels across languages. Thus MultiMASC is far more ambitious and, correspondingly, more labor-intensive collaborative project than the Language Library, but promises to deliver resources that can be used to train learn-

[34] http://www.languagelibrary.eu

ing algorithms and provide new insights about relationships across linguistic levels and languages that the Language Library cannot provide.

The second new collaborative project, the Groningen Meaning Bank (GMB) [1], has established a collaborative effort to provide manual validations of automatically-produced annotations for several linguistic layers from part of speech through discourse structure. Validation is done by volunteer linguists. The data are chosen to be in the public domain; interestingly, the project has chosen the MASC data as a part of its corpus. However, the MASC annotations for phenomena included in the GMB are not used but rather re-generated and hand validated, thus effectively duplicating the work done for MASC. Given that one goal of collaborative annotation is to avoid duplication of effort, it is somewhat tautological for a collaborative project to (in part) discard and re-do the same work as another collaborative effort. The GMB does not use the MASC annotations because of differences in tokenization and (some) annotation categories that are incompatible with their annotation tools.

The GMB has recently established a "game with a purpose" called *Wordrobe* that enables collecting validations from non-experts, which, if enough redundant validations are collected, can provide reliable results by majority vote (see, for example, [22]). The success of *Wordrobe* and *PhraseDetectives*[35] for co-reference annotation (which is also annotating MASC data), together with crowdsourcing in general, suggest the possibility to exploit these strategies for development of ANC-OLI annotations. However, although crowdsourcing dramatically reduces the overhead for gathering validated annotations on a relatively large scale, it is not without some cost for setting up, collecting, evaluating, and preparing the results. The purely collaborative development model of the ANC-OLI requires considerably less investment, since the only requirement is conversion of annotations in different formats to GrAF for compatibility. As time goes on, fewer and fewer annotations are contributed in a format for which a converter has not already developed, if they are not contributed in GrAF itself, thus further reducing the overhead. As a result, of the foreseeable future the ANC-OLI will likely not pursue this development option.

5 Looking Forward

The eventual vision for the ANC-OLI is to expand to include additional resources–not only corpora but also lexicons, lists, etc.–not only in English but also in multiple languages. As mentioned earlier in section 3.2.5, the multi-lingual effort starts with MultiMASC, which will immediately expand MASC and the collaboration effort upon which it depends by exploiting the infrastructure and expertise established in the ANC-OLI to support development in other languages. Although this effort has only been recently launched, it has already drawn substantial interest within the community. Development of MultiMASC will expand the collaborative activity of the ANC-OLI to include the creation of comparable corpora, for which we have pub-

[35] http://anawiki.essex.ac.uk/phrasedetectives/

lished a first set of guidelines, together with an incremental process for developing a fully inter-linked multi-lingual network of linguistic annotations [10].

We envision linkage across hundreds of languages among linguistic phenomena at many levels, e.g., part-of-speech categories, syntactic structures, paraphrases, semantic roles, named entities, events, etc. For example, Figure 1 depicts linkage among several languages for lexical units representing a common semantic role, in this case the EVENT of "buying". Such inter-linkage would utilize a reference set of categories residing in a data category registry such as ISOCat or OLiA that provides information about the annotation content and, more importantly, cross-references linguistic annotations using the same *conceptual* categories, regardless of physical label, within all of the inter-linked resources. Additional linkage to resources such as WordNet and FrameNet, which are themselves linked to wordnets and framenets in other languages, would add another dimension to this resource network, which would in turn enable cross-linguistic and inter-layer studies on a scale that is currently impossible. Ideally, this network would ultimately be available as Linked Data (see Chiarcos, et al., in this volume) so that the technologies supporting the Semantic Web can be exploited for access and search.

6 Conclusion

A community-wide, collaborative effort to produce high quality annotated corpora is one of the very few possible ways to address the high costs of resource production and ensure that the entire community, including large teams as well as individual researchers, has access and means to use these resources in their work. The ANC-OLI represents the first and largest collaborative effort of its kind, and it should provide a model for new resource development projects.

At present, the obstacles to open collaborative efforts are twofold. The most formidable is the requirement for open data, which is limited by established publication practices and, even where openness is promoted, the influence of the default "share-alike" mode of licensing that can limit use and distribution for some segments of the community. The second is the mindset of the community itself, which must be changed so that "giving back" is reflexive, even if it requires additional effort. We do not imagine either of these obstacles will be overcome easily, but at the same time, it is clear that these cultural shifts are underway and inevitable. We hope that once these shifts are complete, the ANC-OLI will be seen as a pioneering project for openness and collaborative development, upon which others have successfully built.

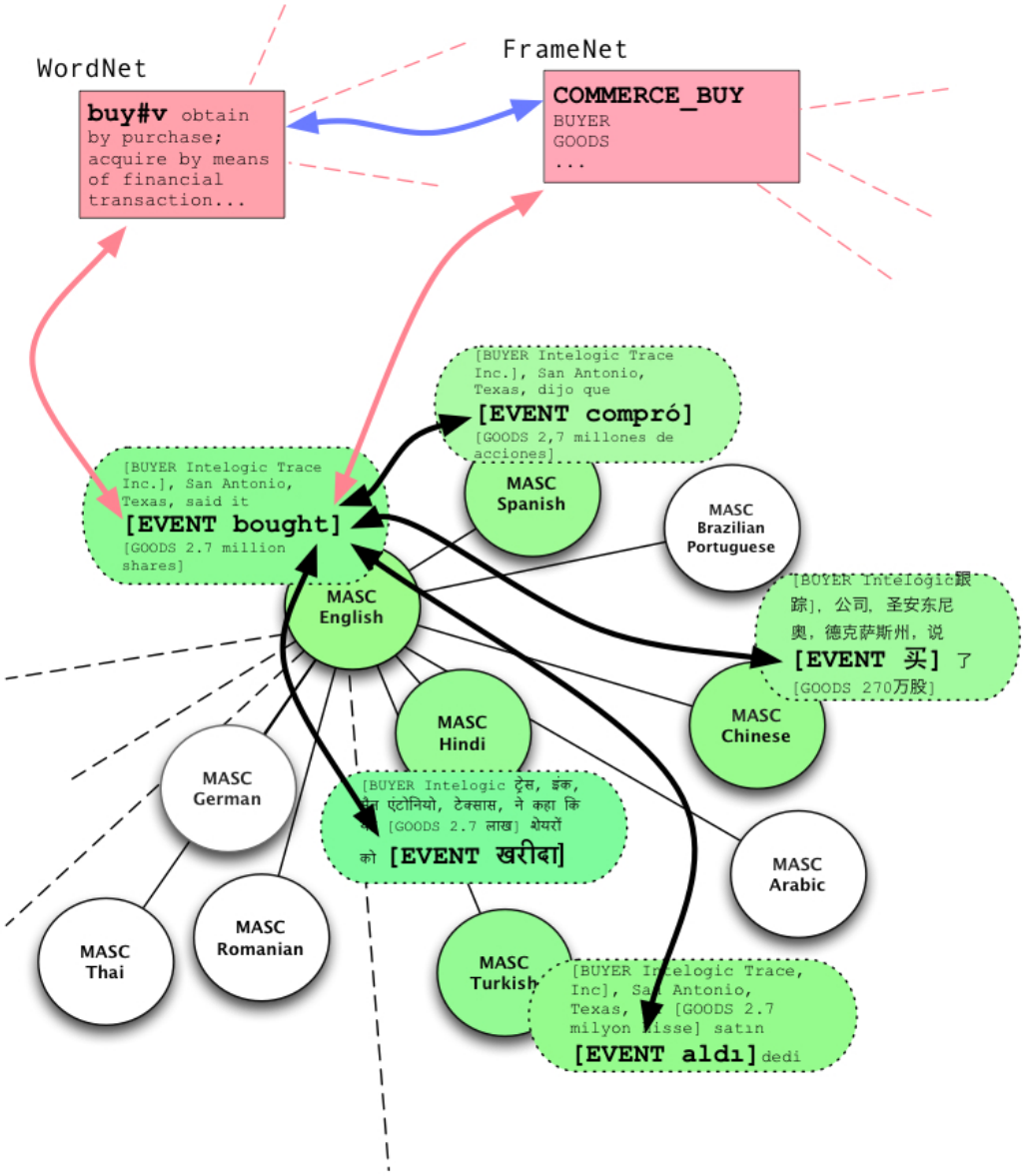

Fig. 1 Overview of MultiMASC

Acknowledgements

This work was supported in part by National Science Foundation grant CRI-0708952.

References

1. Basile, V., Bos, J., Evang, K., Venhuizen, N.: Developing a Large Semantically Annotated Corpus. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 3196–3200 (2012)
2. Bird, S., Liberman, M.: A Formal Framework for Linguistic Annotation. Speech Communication **33**(1-2), 23–60 (2001)
3. Chiarcos, C.: An Ontology of Linguistic Annotations. LDV Forum **23**(1), 1–16 (2008)
4. Chiarcos, C.: Ontologies of Linguistic Annotation: Survey and Perspectives. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) (2012)
5. Chiarcos, C., Ritz, J., Stede, M.: By All These Lovely Tokens... Merging Conflicting Tokenizations. Language Resources and Evaluation **46**(1), 53–74 (2012)
6. Dipper, S.: XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: R. Eckstein, R. Tolksdorf (eds.) Berliner XML Tage, pp. 39–50 (2005)
7. Farrar, S., Langendoen, D.T.: An OWL-DL Implementation of GOLD: An Ontology for the Semantic Web. In: A. Witt, D. Metzing (eds.) Linguistic Modeling of Information and Markup Languages. Springer, Dordrecht (2010)
8. Ferrucci, D., Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering **10**(3-4), 327–348 (2004)
9. Fillmore, C.J., Jurafsky, D., Ide, N., Macleod, C.: An American National Corpus: A Proposal. In: Proceedings of the First Annual Conference on Language Resources and Evaluation, pp. 965–969. European Language Resources Association, Paris (1998)
10. Ide, N.: MultiMASC: An Open Linguistic Infrastructure for Language Research. In: R. Rapp, M. Tadic, S. Sharoof, P. Zweignebaum (eds.) Proceedings of the Fifth Workshop on Building and Using Comparable Corpora. European Language Resources Association (ELRA) (2012)
11. Ide, N., Romary, L.: International Standard for a Linguistic Annotation Framework. Natural Language Engineering **10**(3–4), 211–225 (2004)
12. Ide, N., Suderman, K.: An Open Linguistic Infrastructure for American English. In: Proceedings of the Fifth Language Resources and Evaluation Conference (LREC). European Language Resources Association, Paris (2006)
13. Ide, N., Suderman, K.: GrAF: A Graph-based Format for Linguistic Annotations. In: Proceedings of the First Linguistic Annotation Workshop, pp. 1–8. Prague (2007)
14. Ide, N., Suderman, K.: The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. Language Resources and Evaluation (In Press)
15. Language Resource Management – Linguistic Annotation Framework. International Standard ISO 24612 (2012)
16. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI Meeting Corpus. In: Proceedings of ICASSP-03, pp. 364–367 (2003)
17. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210 (2004)
18. Marc Kemps-Snijders Menzo Windhouwer, P.W., Wright, S.E.: ISOCat: Corralling Data Categories in the Wild. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA) (2008)

19. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993)
20. Miller, S., Guinness, J., Zamanian, A.: Name Tagging with Word Clusters and Discriminative Training. In: D.M. Susan Dumais, S. Roukos (eds.) HLT-NAACL 2004: Main Proceedings , pp. 337–342. Association for Computational Linguistics, Boston, Massachusetts, USA (2004)
21. Miller, S., Guinness, J., Zamanian, A.: Name Tagging with Word Clusters and Discriminative Training. In: Proceedings of Human Language Technologies, pp. 337–342 (2004)
22. Nowak, S., Rüger, S.: How Reliable Are Annotations Via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation. In: Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10, pp. 557–566. ACM, New York, NY, USA (2010)
23. Passonneau, R.J., Baker, C.F., Fellbaum, C., Ide, N.: The MASC Word Sense Corpus. In: N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
24. Pradhan, S.S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: OntoNotes: A Unified Relational Semantic Representation. In: ICSC '07: Proceedings of the International Conference on Semantic Computing, pp. 517–526. IEEE Computer Society, Washington, DC, USA (2007)
25. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF (Working Draft). Tech. rep., W3C (2007). URL http://www.w3.org/TR/2007/WD-rdf-sparql-query-20070326/