Case study: The Manually Annotated Sub-Corpus

Nancy Ide

Abstract

1 Introduction

This case study describes the creation process for the Manually Annotated Sub-Corpus (MASC), which is a subset of the Open American National Corpus (OANC). The OANC is itself a subset of the American National Corpus (ANC). Each of these corpora represents a distinct evolutionary stage in our approach to corpus-building and delivery that reflect adaptations to both changing community needs and advances in best practices for creating and representing linguistically annotated corpora. We therefore describe the procedures involved in producing the ANC and OANC before focusing on MASC, which is the jewel in the crown of corpora produced by the ANC project.

2 Background: The ANC

The ANC was motivated by developers of major linguistic resources such as FrameNet[1] and Nomlex[2], who had been extracting usage examples from the 100 million-word British National Corpus (BNC), the largest corpus of English across several genres that was available at the time. These examples, which served as the basis for developing templates for the description of semantic arguments and the like, were often unusable or misrepresentative due to significant syntactic differ-

Nancy Ide
Vassar College, Poughkeepsie, New York 12604-0520 USA e-mail: ide@cs.vassar.edu

[1] http://www.icsi.berkeley.edu/ framenet

[2] http://nlp.cs.nyu.edu/nomlex/index.html

ences between British and American English. As a result, in 1998 a group of computational linguists proposed the creation of an American counterpart to the BNC, in order to provide examples of contemporary American English usage for computational linguistics research and resource development [8]. With that proposal, the ANC project was born.

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. As for the BNC, funding for the ANC would be sought from publishers who needed American language data for the development of major dictionaries, thesauri, language learning textbooks, et cetera. However, beyond these similarities, the ANC was planned from the outset to differ from the BNC in a few significant ways. First, additional genres would be included, especially those that did not exist when the BNC was published in 1994, such as (we)blogs, chats, and web data in general. The ANC would also include, in addition to the core 100 million words, a varied component of data, which would effectively consist of any additional data we could obtain, in any genre, and of any size. In addition, the ANC would include texts produced only after 1990 so as to reflect contemporary American English usage, and would systematically add a layer of approximately 10 million words of newly produced data every five years.

Another major difference between the two corpora would be the representation of the data and its annotations. The BNC exists as a single enormous SGML (now, XML) document, with hand-validated part of speech annotations included in the internal markup. By the time the ANC was under development, the use of large corpora for computational linguistics research had sky-rocketed, and several preferred representation methods had emergedin particular, stand-off representations for annotations of linguistic data, which were stored separately and pointed to the spans in a text to which they referred, were favored over annotations that were interspersed within the text. The ANC annotations would therefore be represented in stand-off form, so as to allow, for example, multiple annotations of the same type (e.g., part of speech annotations produced by several different systems). Finally, the ANC would include several types of linguistic annotation beyond the part-of-speech annotations in the BNC, including (to begin) automatically produced shallow syntax and named entities.

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. The majority of the fifteen or so publishers who did contribute funding to the ANC included several Japanese publishers of texts on English as a second language and a subset of the same British publishers who had supported the BNC. These publishers, together with a handful of major software developers, provided a base of financial support for the project over a 3-year period, but nothing like the support that had been provided to the BNC. After a time, the ANC project also secured a small grant from the National Science Foundation for ANC development. All in all, the ANC

secured about $400,000 to support its first 4 years, orders of magnitude less than supported development of the BNC.

2.1 Data Acquisition

British publishers provided the bulk of the data in the 100 million-word BNC. The plan for the ANC was that the members of the ANC consortium, which included both publishers and software vendors, would do the same for the ANC. However, only a very few of the ANC consortium members eventually contributed data to the corpus.[3] Some additional data was obtained through contributions from the creators of existing corpora such as the Indiana Center for Intercultural Communication (ICIC) Corpus of Philanthropic Fundraising Discourse[4] and the Charlotte Narrative and Conversation Collection (CNCC)[5]. However, without substantial contributions of data from publishers and other sources, data acquisition became a major issue for development of the ANC.

Over the past several years, computational linguists have turned to the web as a source of language data, and several years ago the proponents of the web-as-corpus predicted that development of corpora like the ANC was a thing of the past. The most common counter-argument in favor of a resource like the ANC is that a web corpus is not representative of general language use; for example, one study showed that web language is highly skewed toward dense, information-packed prose [12], and another recently expounded some of the shortcomings of unedited web data for NLP research [20]. However, the most significant argument against the web-as-corpus is that studies involving web data are not replicable, since the "corpus" and any accompanying annotations cannot be redistributed for use by others. Copyright law, at least in the U.S., specifies that all web data are copyrighted unless explicitly indicated to be in the public domain or licensed to be redistributable through a mechanism such as Creative Commons[6]. Contrary to popular opinion, this includes all of the data in Wikipedia, which has been heavily used in NLP research in recent years.

While the fact that web data is implicitly copyrighted provides some justification for development of a resource like the ANC, this fact also presented the greatest obstacle to ANC data acquisition. Data on the web—including PDF and other documents that are not typically included in web corpora–are the most likely source of material for inclusion in the ANC; however, the vast majority of web data in the public domain is at least 50 years old because of copyright expiration, and the ANC requires data produced since 1990. The search for more recent web documents that are explicitly in the public domain or licensed for unrestricted reuse is therefore not

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

[4] http://liberalarts.iupui.edu/icic/research/corpus_of_philanthropic_fundraising_discourse

[5] http://nsv.uncc.edu/nsv/narratives

[6] http://creativecommons.org/

only time-consuming, but also yields relatively meager results. As a result, the ANC had to rely primarily on government sites for public domain documents, as well as web archives of technical documents such as Biomed[7] and the Public Library of Science[8]. To attempt to gather data from other sources, the ANC project put up a web interface[9] to enable contributions of texts from donors such as college students, who are asked to contribute the essays, fiction, etc. they have written for classes; an ANC Facebook page is maintained to reach out to college students for contributions.[10]

2.2 Data Preparation

ANC data was obtained from a variety of sources and came in many different formats, including plain text, HTML, Word doc and RTF format, PDF, and various publishing software formats such as Quark Express. The most significant effort in the early stages of the project was therefore to transform the data into a format suitable for annotation. Depending on the original format, this demanded a more or less complex series of steps. Unexpectedly, some of the the easiest formats to preprocess turned out to be Word doc and RTF; after a document in one of these formats is opened in Open Office[11], it can be exported in TEI XML using an available plug-in, transformed to XML Corpus Encoding Standard (XCES) [11] format using an XSLT style sheet, and finally loaded into GATE, which separates textual content from XML markup and preserves the markup in standoff form. Thus for a Word or RTF document, the full processing pipeline is a push-button operation. Documents already encoded in XML, such as those obtained from PLoS and Biomed, require only the last few steps. However, other formats proved to be more problematic. Text cannot be straightforwardly extracted from PDF documents without requiring semi-automatic post-editing to eliminate page numbers and running heads, etc., and we have so far discovered no method to extract text from multi-column PDF that does not require prohibitively extensive post-editing.[12] Formats such as Quark Express, which are used to represent print-ready documents for publishers, present problems with special characters such as ligatures, initial capitals, etc. HTML poses its own set of well-known problems, which are documented in detail in proceedings of the CLEANEVAL exercises[13]. Plain text is easy to process, but lacks formatting information beyond the identification of paragraph boundaries.

---

[7] http://www.biomedcentral.com/

[8] http://www.plos.org

[9] http://www.anc.org/contribute/texts/

[10] To date, we have collected over five million words of college essays and fiction contributed by college students.

[11] http://www.openoffice.org

[12] For this reason, we were unable to include a million words of contributed data from the ACL Anthology in the ANC.

[13] http://cleaneval.sigwac.org.uk/

In addition to the formats mentioned above, the ANC often received data rendered in an arbitrary XML format that provides some kind of annotation. While this would seem to be ideal since XSLT could be used to transform it to XCES, it should never be assumed that one person's XML is mappable to another's. For example, the ICSI Meeting Corpus[14], consisting of spoken transcripts of multi-participant meetings, was contributed to the ANC in an XML format that encloses every distinct fragment of the transcript within a `<Segment>` element, including not only spans of speech, but also"events" such as microphone noise, laughing, etc., and added information such as comments by the transcribers. Because there is no embedding of `<Segment>` elements in the transcripts, extensive processing is required to rejoin parts of a speaker turn that are separated by a segment indicating an interruption (noise, etc.) or transcriber comment. Because these interruptions frequently occur in mid-sentence, the separation poses problems for subsequent part of speech and syntactic analysis. It is, however, often cost-prohibitive to render contributed annotations in an optimal form, and in such cases the data and annotations were released "as is".

The ANC project committed itself from the start to using state-of-the-art standards and best practices, and to make the corpus as widely usable as possible. In the First Release, in order to allow for maximum flexibility, the ANC data used a UTF-16 character encoding[15], which can represent a very wide range of characters. This turned out to be more cumbersome than helpful, given that many software systems do not support UTF-16, and those that do often require special processing. Therefore, all ANC data used a UTF-8 character encoding from the Second Release onward.

2.3 Annotation

Annotation of the ANC data was accomplished primarily with the General Architecture for Text Engineering (GATE) system developed by the University of Sheffield [5]. GATE implements a pipeline architecture for annotating corpora by allowing for the application of a series of software components. GATE provides Java software plugins for a variety of corpus annotation tasks such as part of speech tagging, several kinds of syntactic analysis, named entity recognition, and co-reference resolution, as well as a machine learning module and sophisticated mechanisms for ontology development and use. The feature of primary value to the ANC project is the ability to add or replace modules in the pipeline for processing specific to our needs. The ANC project developed GATE plugins for ANC-specific processing and a Java-based scripting language that enables us to pipeline texts through a series of annotation tools for sentence splitting, tokenization, lemmatization, part of speech

[14] http://www1.icsi.berkeley.edu/Speech/mr/

[15] Defined in ISO/IEC 10646.

annotation, noun and verb phrase chunking, and output the primary and stand-off documents in the final ANC format.[16]

The ANC project had funding to cover only spot-checking of the annotations produced using GATE modules. However, several nite state transducers were implemented using the Java Annotation Patterns Engine (JAPE)[17] to massage the output of built-in GATE modules, based on analysis of systematic errors observed in the output.[18]

2.4 Format

Development of the BNC included the production of a software system for searching the corpus, generating concordances, etc. (XIARA). It was clear from the outset that without the similar funding, the ANC project would be unable to produce search and query software for the ANC. The alternative was to represent the corpus and its annotations in such a way that it could be used with existing software, including XIARA, widely used commercial concordance software (e.g., MonoConc, WordSmith), and text engineering systems that existed at the time such as GATE[19]. This meant that the ANC and its annotations had to be represented in a format that could be straightforwardly transduced to virtually any other input format required by such softwarea non-trivial requirement. In addition, the layering of annotations in the ANC and the inclusion multiple annotations of the same type dictated the use of a stand-off annotation representation format. In a stand-off representation, annotations reside in a separate document or documents linked to the primary data, and the primary data remains "read-only".

At the time the ANC project was begun, members of the project were involved in development of a stand-off format for linguistically-annotated data under development by the International Standards Organization (ISO) TC37 SC4 (Language Resource Management), which was intended to reflect the state-of-the-art. Therefore, it was decided that the ANC would serve as the poster child for the ISO group's Linguistic Annotation Framework (LAF) [13, 14], which provides a general framework for representing annotations serialized in XML by the Graph Annotation Format (GrAF) [16]. GrAF's underlying data model comprises an acyclic di-graph decorated with feature structures (coupled with a moderate admixture of algebra, e.g. disjunction, sets), grounded in n-dimensional regions of primary data (see [17] for a full description of LAF and its GrAF XML serialization). The graph

[16] The ANC maintains a GATE plugin repository, which includes import and export modules for annotated documents in GrAF (see Section 2.4), at http://www.anc.org/tools/gate/gate-update-site.xml.

[17] http://gate.ac.uk/sale/tao/splitch8.html

[18] Some of these modules were developed or improved by students at Vassar College, who did the analysis and JAPE rule-writing as a term project for an advanced undergraduate course on Computational Linguistics.

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

itself is a generalization of models for a wide range of phenomena, including syntax trees, semantic networks, W3Cs RDF/OWL, the Unified Modeling Language (UML), entity-relation (ER) models for databases, etc.–not to mention the overall structure of the web, which is a dense inter-connected network of objects–and feature structures have long been used to represent both simple and complex linguistic annotations. Because of the generality of the underlying data model, GrAF is trivially mappable to many existing and evolving formats, and the rendering of ANC data and annotations in GrAF thus satisfied a primary criterion for ANC design: the ability to transduce ANC data and annotations into formats required by various software systems.

The development of LAF/GrAF and the ANC was symbiotic: the ANC served as a testing ground for LAF, which in turn evolved based on the experience gained in representing the ANC. This meant that the representation format for the ANC changed from release to release. The stand-off version of the First Release was represented using a very early format that resembled GrAF only in terms of structure; the Second Release used an early version GrAF, which changed slightly over the following years but is trivially transformed to the final version, published in 2012 [19]. To facilitate use of the ANC, the Second Release included a first version of the ANC Tool, which generates parts or all of the corpus with user-selected annotations any of several formats usable by UIMA, NLTK, XAIRA, MonoConcPro and WordSmith, as well as CoNLL format and inline XML.The ANC Tool subsequently evolved into a suite of GrAF APIs, together with a web service that provides transduction from GrAF to an increasing number of formats as well as easy means to develop transducers to other formats, described in detail below in Section 5.1.

2.5 Distribution and Delivery

In 2003, the ANC produced its first release of 11 million words of data, which included a wide range of genres of both spoken and written data.[20] Annotations included word and sentence boundaries and part-of-speech annotation produced by two different taggers: the "Hepple tagger", which uses the Penn part of speech tags, and the "Biber tagger", which uses a superset of the CLAWS part of speech tags used to tag the BNC. The annotations in this release were represented in standoff form–that is, annotations were not included inline with the text but rather provided as separate files with links into the data. To our knowledge, the ANC First Release was the first large, publicly available corpus to be published with standoff annotations. Because of the lack of software for handling standoff annotations at the time, a version of the ANC First Release with inline annotations was also included in the distribution.

In 2005, the ANC released an additional 11 million words, bringing the size of the ANC to 22 million words. The Second Release includes data from additional

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

genres, most notably a sizable sub-corpus of blog data, biomedical and technical reports, and the 9/11 Report issued 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) and two additional part of speech annotations using the CLAWS 5 and 7 tags to enable comparison with BNC data. Both the First and Second Releases of the ANC are distributed through the Linguistic Data Consortium (LDC) for a reproduction fee of $75.00 for non-members who will use it for research purposes only. Frequency data for the corpus, including pos frequency data, bigrams, etc., is available on the ANC website. After 2005, the ANC project had no more funding, and production of additional data came to a halt.

3 Open ANC

In 2006, the ANC project made 15 million of the ANCs 22 million words, called "the Open ANC" (OANC), available from the ANC website. Although the OANC is not as broadly representative as the BNC, it nonetheless contains the widest variety of genres–including contemporary genres such as blogs, email, etc.–of any large, redistributable corpus of contemporary English in existence. Most notably, the OANC subset of the ANC is free of licensing restrictions, and therefore is available for download to anyone for any purpose, research or commercial. The OANC distribution model of completely open access is a step beyond licenses such as Creative Common "share-alike" and the GNU Public License, which require redistribution under the same license and are therefore prohibitive for commercial users. At the same time, acquisition of fully open data can be a very difficult and time-consuming process, either because of the necessity to search for web materials clearly labeled as public domain or issued under a license like Creative Commons Attribution (CC-BY), or the effort involved in obtaining permission from authors to distribute their data with no constraints. In 2006, the OANC was a pioneer in the move toward open linguistically-annotated data; since then, the community has begun to actively embrace the idea of fully free and open access to resources, seen for example in the recent creation of the Linguistic Linked Open Data (LLOD) cloud[21] [3].

The OANC was publicized as a community-based project, with the expectation that with freely available data, members of the community would contribute annotations for use by others. Several contributions were received, including BBN Named Entities (inline format), syntactic parses in various formats, coreference (anaphora) annotations of Slate journal articles, and CLAWS 5 and 7 part of speech tags for the ANC First Release data.[22] Satoshi Sekine also contributed an *n*-gram search engine for the OANC.[23]. However, the contributions that were received fell far short of

[21] http://linguistics.okfn.org/resources/llod/

[22] Available at http://www.anc.org/data/oanc/contributed-annotations/

[23] http://www.anc.org/data/oanc/ngram/

our expectations. The experiment to create an "Open Linguistic Infrastructure" for American English [15], which would include contributed annotations at all linguistic levels, link semantic annotations of ANC data to databases such as WordNet and FrameNet, provide derived data and other resources, etc. did not become a reality until the development of MASC, as described in the next section.

4 MASC

In 2007 the ANC received a substantial grant from the U.S. National Science Foundation[24] to produce a Manually Annotated Sub-Corpus (MASC) of the ANC. The grant was awarded on the basis of a mandate from the US Computational Linguistics community to create a high-quality gold standard corpus that includes a broad and representative range of genres.[25] The demand for a broad genre corpus was a reaction to the domain-specificity of available corpora with multiple layers of annotation, which included the one million word Wall Street Journal corpus known as the Penn Treebank [23] and the OntoNotes corpus of newswire, broadcast news, and broadcast conversation[26]. The NSF grant provided no funding to add data to the existing OANC, but rather provided funds to validate automatically-produced annotations for part of speech, shallow parse, and named entities, and to manually add annotations for WordNet senses and FrameNet frames to portions of the corpus. Partners in the project included the FrameNet team at ICSI, UC Berkeley; the WordNet team at Princeton; and researchers at Columbia University.

4.1 The Data

MASC is a 500,000 word corpus of post-1990s American English comprised of texts from nineteen genres of spoken and written language data in roughly equal amounts, shown in Figure 1). The data were drawn primarily from the OANC, described above in Section 3, but to provide additional genres and, especially, to ensure that MASC included recent social media data, some texts were drawn from the roughly 50 million words of unrestricted data that was collected for the OANC but never processed due to lack of funding. Roughly 15% of the corpus consists of spoken transcripts, both formal (court and debate) and informal (face-to-face, telephone conversation, etc.); the remaining 85% covers a wide range of written genres, including social media (tweets, blogs).

Where licensing permitted, data for inclusion in MASC was drawn from sources that have already been heavily annotated by others. MASC includes a roughly 50K subset of OANC data that had been previously annotated for PropBank predicate

[24] NSF CRI 0708952

[25] See http://www.anc.org/MASC/About_files/NSF_report-final.pdf

argument structures, Pittsburgh Opinion annotation (opinions, evaluations, sentiments, etc.), and several other linguistic phenomena. MASC also includes a set of small texts from the so-called Language Understanding (LU) Corpus[26] that has been annotated by multiple groups for a wide variety of phenomena, including events and committed belief, plus 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. Most of the annotations of these data have been contributed for inclusion in MASC.

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

Fig. 1 Genre distribution in MASC

The choice of genres to include in MASC was somewhat dependent upon availability of data unrestricted by licensing concerns, but an effort was made to include a range of genres somewhat similar in scope to the BNC, and to include fiction and social media such as blogs, email, and tweets. The corpus also includes government documents, biomedical articles, movie scripts, jokes, and college essays (contributed through our website, as described in Section 2.1), as well as "ficlets" (story fragments to which "prequels" or "sequels" are added by online participants) and Berlitz Travel Guides. An original list of twenty MASC genres included poetry, but it was not possible to find a large enough sample to include in the corpus.

All MASC data was prepared using established procedures and software developed to produce the ANC (See Section 2.2). Because the corpus is small, the data

[26] MASC includes about 5K of the 10K LU corpus, eliminating non-English and translated texts as well as texts that are not free of usage and redistribution restrictions. See https://catalog.ldc.upenn.edu/LDC2009T10.

were checked by hand more thoroughly than much of the ANC data, in order to fix bad line breaks, eliminate spurious or odd characters, etc.

4.2 Annotation

The premise behind MASC from the outset was to provide appropriate data and annotations to serve as the base for a community-wide annotation effort, together with an infrastructure that enables the representation of internally-produced and contributed annotations in a single, usable format that can then be analyzed as it is or ported to any of a variety of other formats, thus enabling its immediate use with common annotation platforms as well as off-the-shelf concordance and analysis software. The aim was to offset some of the high costs of producing high quality linguistic annotations via a distribution of effort, and to solve some of the usability problems for annotations produced at different sites by harmonizing their representation formats.

The annotation types and coverage developed by the MASC project and distributed in the current version of the corpus (3.1) are given in Figure 2[27].

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	506659
Coreference	506659
Discourse structure	506659
Opinion	51243
TimeBank	*55599
PropBank	88530
Committed Belief	4614
Event	4614
Dependency treebank	5434

Fig. 2 Summary of MASC annotations

Annotations for logical structure (titles, headings, sections, etc. down to the level of paragraph), tokens, sentence, part-of-speech and lemma, noun chunks, verb chunks, named entities (Person, Organization, Date, and Location, with subtype information), and coreference were initially generated using built-in GATE mod-

[27] The list does not include WordNet sense annotations because they are not applied to full texts.

ules, most of which belong to GATE's ANNIE suite of tools[28]. The automatically-produced annotations were then checked and corrected manually by undergraduate annotators at Vassar College[29], using the GATE environment. With the exception of coreference and discourse structure, which were added later in the project (see **??**), the procedure was as follows:

1. a gold standard annotation for 10K words of data was created, starting from the automatically-generated annotations by an expert;
2. annotators attended a training session, where they were introduced to the annotation guidelines for the relevant phenomenon and performed sample exercises;
3. at least two, and as many as four, annotators independently corrected the automatically-generated annotations;
4. the corrected versions were compared to the gold standard using GATE's "AnnotationDiff" tool;
5. the corrected versions were compared to each other using GATE's "AnnotationDiff" tool;
6. systematic errors in the automatically-generated data were identified by hand;
7. systematic inconsistencies between annotators were identified by hand.

On the basis of (5), we developed post-processing scripts using GATE's Java Annotation Patterns Engine (JAPE), which provides finite state transduction over annotations based on regular expressions, to automatically correct systematic errors. In some cases, issues were addressed by adding to the gazetteer lists (which had already been augmented from the default ANNIE Gazetteer in an earlier project) and/or modifying or adding to the lexicon used in the part-of-speech tagger. Results from (6) were used to improve the annotation guidelines provided to student annotators.[30]

We applied GATE's Performance Evaluation tools[31] to provide basic statistics, including precision, recall, and f-score, which enabled us to identify the annotations types that were most reliably corrected by annotators and those that posed more difficulties. Among the non-controversial annotation types, we found that noun chunks were most reliably identified by annotators, and apart from part-of-speech (which was handled separately–see below), verb chunks posed the most difficulties. We encountered the well-known problem of ambiguity in determining named entities for locations and organizations, which was addressed by adding an attribute *locOrgAmbig* to *Organization* and *Location* annotations–this was, however, introduced later in the project and therefore not applied systematically.

The automatically-generated annotations were then post-processed by adding the JAPE scripts, updated gazetteer, and lexicon into the pipeline to correct systematic

[28] http://gate.ac.uk/sale/tao/splitch6.html#x9-1260006

[29] Primarily, the students were Cognitive Science majors with a Linguistics emphasis. Over the four years of the project, sixteen different students worked on validation.

[30] All of the MASC project's annotation guidelines are accessible from http://www.anc.org/wiki/#AnnotationValidation.

[31] http://gate.ac.uk/sale/tao/splitch10.html

errors, and the newly-generated annotations were given to the annotators for correction; however, in this phase, the newly-generated annotations were first given to a single annotator, followed by a second annotator who reviewed and corrected (where necessary) the first annotator's work. As might be expected, as students worked on the manual correction it became clear that some were more proficient than others, and this was taken into account when assigning students to work with documents. Depending on the difficulty of the annotation type and the quality expected from the previous annotator(s), a third annotator might be assigned to review and correct the second annotator's results.

MASC annotations can be separated into two types:

1. annotations for "non-controversial" phenomena, which for in the MASC project includes logical structure, tokens, sentences, noun chunks, verb chunks, and (most) named entities; while there may be differences in interpretation of these phenomena across different annotation projects, these annotation types can be consistently and unambiguously annotated to conform to our annotation guidelines such that there is always an identifiable, correct annotation.
2. annotations where there may exist legitimate alternative annotations, even given clear guidelines; these annotations include part-of-speech, co-reference, and discourse structure.[32]

Annotations of type (2) were treated slightly differently from the others. Tokenization and part-of-speech annotation for the entire corpus was initially produced with GATE's ANNIE POS tagger, which uses the Penn tag set. Shortly after the beginning of the MASC project, the Penn Treebank (PTB) project was contracted to provide PTB syntax annotations over the entire corpus, which would include hand-validated part-of-speech tags also using the Penn tag set. Because part-of-speech validation is a difficult task for annotators without relatively sophisticated linguistic background and/or training, we did not make a substantial investment in hand-correcting the POS tags produced by the ANNIE tagger. Instead, we performed the same kind of analysis for systematic errors as we had down for other annotation types and created JAPE scripts to correct systematic errors. This task was made easier because the scripts could reference annotations for noun and verb chunks and entities to locate erroneous tag assignments. For example, in a phrase such as "the winding road", "winding" may be tagged as a present participle verb (VBG); the scripts would specify that the VBG tag should be changed to adjective (JJ) when it is associated with a word that appears prior to a noun *and* falls within a span annotated as a noun chunk, etc.

Our intention was to use the hand-corrected part-of-speech tags produced by the Penn Treebank (PTB) project to correct remaining erroneous tags produced by GATE's ANNIE tagger. Both sets of tags would be retained; we noted that there were some tagging differences between the PTB and ANNIE tags that represented different tagging philosophies rather than errors as such–for example, a word like "please" in the phrase "please, help..." would be tagged by ANNIE as a verb (VB),

[32] Sense and frame element annotations were handled separately; see Chapter XX in this volume.

while PTB would tag it as an interjection (UH). Other differences resulted from variant tokenizations, most notable the handling of hyphenated words such as ”oscar-winning”, which the ANNIE tokenizer[33] treats as one token and PTB treats as three (”oscar”, ”-”, and ”winning”). Because of the tokenization, the PTB part-of-speech assignment is `oscar/NNP, -/HYPH, winning/VBG`, whereas ANNIE tags ”oscar-winning” as an adjective. We decided which tag to retain in the ANNIE part-of-speech tagging on a case-by-case basis; the most notable departure from the PTB tagging is the tokenization and tagging of hyphenated adjectives, as shown above. [34]

The PTB annotations were received in the inline, bracketed format used for the Penn Treebank syntactic annotations available through the LDC[35], which demanded development of a converter to extract the inline annotations and align them with the original MASC text. The task of comparing the PTB part-of-speech tags to the ANNIE tags therefore required a non-trivial alignment exercise to determine corresponding tokens between the two parer-of-speech annotations, followed by creation of a mapping between the two taggings to indicate the circumstances under which which the ANNIE tags were to be changed. We developed a small web application[36] that shows the tokenization and part-of-speech assignment for each word in each document in MASC, with differences highlighted in a different colors depending on the type. This enabled us to readily identify the variations and decide on a mapping from the PTB tags to the ANNIE tags where needed. Although we encountered occasional errors in the PTB tagging, these were left as we had received them from the PTB project.[37]

Annotations for co-reference and discourse structure were added to MASC after the initial release of the full corpus, and for both we continued the strategy of passing automatically-produced annotations to annotators in sequence, such that the second annotator worked with the prior annotator’s results. For each of these annotation types, three annotators were assigned. However, in this case, annotators received additional training, and an annotator would change a previous annotator’s annotation only if it was clearly in error or had not been done at all. However, when the difference could potentially be attributed to a difference of opinion, the annotation was not changed, but rather, a new annotation was added as a feature and the responsible annotator was identified. Annotators also provided a confidence level for each annotation (the default was “high”, so confidence level was explicitly noted

[33] We created a post-processing JAPE script that modifies the default ANNIE tokenization slightly.

[34] The PTB project changed its tokenization, which originally did not break hyphenated words, because of difficulties with cases such as “New York-based” encountered in the Unified Linguistic Annotation project (see https://catalog.ldc.upenn.edu/LDC2009T07). However, this disallowed tagging the hyphenated word as an adjective, which, despite the need to manually correct tokenizations such as `New+York-based`, was deemed preferable.

[35] https://catalog.ldc.upenn.edu/LDC99T42

[36] http://anc-projects.appspot.com/ptbpennposcompare

[37] Because of the unexpected difficulty of correcting the ANNIE tags by this method, the first release of the full MASC (version 3.0.0) did not contain the tags corrected from the PTB data, but had been pos-processed with JAPE scripts to correct systematic errors.

only in more dubious cases). Cases where there had been a difference of opinion were examined by two experts after all three annotators had completed their work, and a primary annotation was selected. Where there was considerable ambiguity, a feature was included on the annotation reflecting the alternative interpretation.

Co-reference annotations were generated by applying GATE's ANNIE Nominal and Pronominal co-referencers to our previously validated noun chunks and named entities. Annotations for discourse structure were produced by applying a discourse parser developed at Universitatea Alexandru Ioan Cuza in Romania. The annotations produced by this tool include clause boundaries and discourse markers; a feature indicating the nucleus/satellite relations among clauses (as specified in Rhetorial Structure Theory [22]) was manually added to each of the clause annotations.

A focus of the MASC project was to provide corpus evidence to support an effort to harmonize sense distinctions in WordNet and FrameNet [1, 6]. For this purpose, the MASC project also produced annotations over portions of the corpus for WordNet senses and FrameNet frames and frame elements. The procedures for sense and frame element annotation differed significantly from those for other MASC annotations, involving substantial inter-annotators agreement studies in the case of sense annotation and intensive manual annotation for FrameNet frame elements. The strategies for sense and frame annotation are fully described in Chapter XX, "The MASC Sentence Corpus", in this volume. Full text annotations for FrameNet frame elements were also produced for approximately 40K words of MASC data in addition to the annotation of individual sentence described in Chapter XX, following the same procedures outlined there.

4.3 Contributed Annotations

From the outset, MASC was intended to be a community-based project, with MASC serving as the basis for community-contributed annotations and, ultimately, an "open linguistic infrastructure" for linguistically-annotated data as described in [10]. Contributed annotations are transduced to GrAF by the ANC project, so that all MASC annotations are in a common format in order to be usable together. So far, the following annotations have been contributed:

- MASC-NEWS[38] automatic annotation of MASC for named entities and word senses based on BabelNet[39] [24].
- A lexical substitution corpus CoInCo (Concepts in Context) based on contiguous texts from MASC, which contains substitute words collected via crowdsourcing for every content word in selected (complete) text files [21].
- MASC AMT Word Sense Annotation[40], 1000 occurrences of each of 45 words for WordNet 3.1 word senses drawn from MASC, including a mix of nouns (n),

[38] http://lcl.uniroma1.it/MASC-NEWS/

[39] http://babelnet.org/

[40] http://dx.doi.org/10.7916/D80V89XH

verbs (v), and adjectives (j) from texts in a range of genres. Each word was labeled by approximately 25 different annotators, for a total of roughly 1M total annotations.

5 Format

Like the OANC, MASC is represented in GrAF. GrAF represents stand-off annotations by containing each annotation layer in a separate XML document linked to the primary data. Each text in the corpus is provided in UTF-8 character encoding in a separate file, which includes no annotation or markup of any kind. Each text is associated with a set of GrAF standoff files, one for each annotation type, containing the annotations for that text. Each text is also associated with a header document that provides appropriate metadata together with machine-processable information about associated annotations and inter-relations among the annotation layers; and a segmentation of the primary data into minimal regions, which enables the definition of different tokenizations over the text. A resource header provides meta-data for the entire corpus by establishing resource-wide definitions and relations among files, datatypes, and annotations that can enable automatic validation of the resource file structure and contents.[41]

One of the fundamental design criteria for GrAF, especially as opposed to earlier formats such as Annotation Graphs, was to allow for a graph of annotations over the data, where regions of primary data comprise the leaves (terminals) of the graph, and the graph is built up by first associating annotations with those regions and then effectively "layering" annotations by associating them with annotations at lower levels. In MASC, ANNIE tokens are defined over the minimal regions of a text, and noun chunks, verb chunks, named entities, and discourse clauses are linked to the tokens that comprise them. Coreference annotations are linked to named entities or, where no entity is present, noun chunks (in cases where no entity or noun chunk exists, a new annotation *Markable* was introduced and linked to the relevant tokens).

MASC annotations for Penn Treebank syntax and FrameNet were created at different sites and therefore came with their own tokenizations, and in their own format. Conversion from other formats to GrAF can be a non-trivial process. Inline annotations must be extracted from the text and rendered as stand-off, with links into the data; acritical point is that in order to make the annotations compatible (mergeable) with existing MASC annotations, the alignment must be to the "read-only" version of the text that is a part of MASC. Therefore, the process involves not only extraction of the annotations, but also alignment of the annotated text with the appropriate regions of the MASC text in order to determine its location and specify offsets. This process is made especially difficult due to changes in the original text, most notably removal or addition of spaces, line breaks, etc., but also modification of the original text to correct errors or render the text in a form that is more easily

[41] For comprehensive overview of GrAF and its headers, see [17].

processed by available software, for example, rendering all characters in lower case, inserting additional blanks where tokenization is desired, etc.

By far the greatest problem we have encountered in attempting to make all MASC annotations fully compatible results from differences in tokenization. As noted above, GrAF attempts to address the problem by creating a *base segmentation* over the data that chops the text into minimal regions, such that any tokenization–and multiple conflicting tokenizations–can be defined over it. Thus when one scheme tokenizes "can't" as *ca + n't* and another tokenizes it as *can + 't*, the fact that they cover the same span can be detected, which (in principle) makes merging and comparison easier. However, when processing annotations received with their own tokenization, no effort was made to harmonize that tokenization with the ANNIE tokenization that is the basis of most of MASC's annotations. As a result, at present MASC includes three different tokenizations: the ANNIE tokenization, which is the basis for most MASC annotations; the PTB tokenization, which is the basis for the Penn Treebank syntax annotations; and the FrameNet tokenization, which is the basis of the FrameNet annotations.

5.1 Distribution and Delivery

MASC fully open and freely available for any use. The corpus is downloadable from http://www.anc.org/data/masc; it is also available from the Linguistic Data Consortium (LDC)[42]. The full MASC download contains all the MASC texts and annotations in GrAF format and the resource header. Contributed annotations are also included in their original format, where available.

The ANC project provides an API for GrAF that can be used to access and manipulate GrAF annotations directly from Java programs. An independent effort within the European project CLARIN[43] developed a Python implementation of GrAF[44] and an API for mapping data formats used in language documentation into GrAF and back[45] [2]. The GrAF Java API includes a graph renderer that transducer GrAF annotations to the input format for the open source GraphViz graph visualization application[46] to enable visualization of the graphs. More recently, researchers at Universität Potsdam developed a GrAF importer for ANNIS[47] [**?**], which provides powerful annotation query and visualization capabilities.[48]

The ANC project also provides plugins for GATE to input and/or output annotations in GrAF format; a "CAS Consumer" to enable using GrAF annotations in

[42] See https://catalog.ldc.upenn.edu/LDC2013T12

[43] http://www.clarin.eu

[44] Available from https://pypi.python.org/pypi/graf-python/0.3.0

[45] https://poio-api.readthedocs.org/en/latest/

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

[47] http://www.sfb632.uni-potsdam.de/annis/

[48] The ANNIS implementation for accessing MASC annotations is available from http://www.anc.org/software/annis.

the Unstructured Information Management Architecture (UIMA) [7]; and a corpus reader for importing MASC data and annotations into the Natural Language Toolkit (NLTK)[49].

An important delivery mechanism for MASC is ANC2Go [18], a web application that comprises a suite of web services for transducing annotations in GrAF to a variety of other formats. ANC2Go allows the user to create a "customized corpus" by choosing from among available texts and annotations in either of MASC or the OANC, and receive the output in any of a variety of formats. At the present time, the available formats include the following:

- inline XML (suitable for input to XML-aware software);
- token+part of speech (with choice of separation character), a common input format for general-purpose concordance software and numerous parsers;
- word/pos output in a format readable with the NLTK's TaggedCorpusReader;
- CONLL IOB format, used in the Conference on Natural Language Learning[50] shared tasks;
- UIMA CAS, for input to UIMA;
- the W3C Resource Description Framework (RDF).

MASC is currently being imported into the LLOD cloud, and, by virtue of its WordNet and FrameNet annotations, its sense and frame element annotations will be linked to the LLOD instantiations of WordNet and FrameNet.

All tools produced by the ANC project are available for download at http://www.anc.org/software.

6 Retrospect

MASC was–and is–an ambitious project, especially at the time it was begun. It was one of the first corpora to be published with stand-off annotations[51], and was intended to be a poster child for LAF/GrAF, which at the time was at the forefront of state-of-the-art strategies for representing linguistically-annotated data. Together with the OANC, it is also one of the first corpora to be released as a fully open resource, and was an early example of a community-based effort to develop and enhance resources for universal use. As such, development of MASC was a pioneering effort, and its format and distribution model have had a noticeable impact on resource development and distribution in recent years.

Using MASC as a means to both inform GrAF's design and serve as a model of its use was a risk. The aim for MASC was to provide a corpus that would be maximally usable and reusable, and in particular could be used with a wide variety of corpus query, access, and manipulation software. it was also necessary to enable others

[49] http://nltk.org

[50] http://ifarm.nl/signll/conll/

[51] Note that GrAF is a "true" standoff format, as opposed to hybrid standoff formats as described in Chapter III in this volume.

to easily add and contribute annotations. These requirements precluded a solution such as the one adopted in OntoNotes, which developed a special internal format and provides a query and access framework. Our solution was to adopt a representation that was general enough to be transducable to input formats for common tools such as GATE, UIMA, NLTK, and XIARA. As with almost any standard in the field, LAF/GrAF has seen fairly wide adoption, but neither it nor any other format has yet to provide the ultimate standard for representing linguistic annotations. LAF/GrAF's final claim to fame is likely to be its significant influence on the representation of linguistically-annotated data by introducing the graph-based model and its use to enable trivial mapping among formats for interoperability among resources, which now dominates the field.

Our procedure for correcting the core MASC annotations was somewhat unorthodox in that annotators made sequential passes in which they corrected errors or omissions that remained at that point, rather than the parallel annotation strategy typically used in annotation projects. The motivation for this was two-fold. First, we simply did not have the resources for an extensive annotation effort involving annotators with sophisticated linguistic training, but had to rely on (bright) undergraduates with some linguistic sensitivity. Second, the bulk of our annotations are of the non-controversial variety, as described above in Section 4.2, rather than annotations for phenomena such as sense and frame element annotation, where the need for linguistic training and inter-annotator agreement studies is more critical. The exceptions are the later addition of annotations for co-reference and discourse structure, to which we applied the same sequential procedure but with more annotators per document and more systematic documentation of changes. We believe that the quality of the annotations is as high or higher as for comparable annotation efforts such as OntoNotes (see Chapter XX in this volume). We also anticipate that users and contributors will submit error reports that will allow us to correct any remaining errors.

Based on the experience with MASC, we believe that the primary unsolved problem for representing linguistically-annotated corpora is variations in tokenization.[52] It is somewhat ironic that such a low-level linguistic issue so profoundly inhibits the combined use of annotations of the same data produced at different sites. If, as we anticipate, annotation efforts become more community-based–through contributions to existing corpora such as MASC, crowdsourcing, and/or distributed effort involving the human-in-the-loop–this issue will demand a resolution, but to date very little effort has been made to provide one. GrAF's solution of basing all tokenizations on a common base segmentation goes in the right direction, but even if common spans are automatically detected, there is no guarantee that a meaningful resolution of conflicts that would allow for seamless annotation merging will exist in every case. This remains an open problem for the field.

[52] The inability to harmonize the annotations produced for the Language Understanding (LU) corpus provides a notorious example; see also [4] and [9] for related examples and discussion.

7 Conclusion

MASC is the most richly annotated corpus of English available for unrestricted use. The ANC project is currently adding an additional 500K words to the corpus to bring it to one million words; although funding is limited, we can apply our processing and automatic-correction pipeline to annotate the data for (at least) the core MASC annotations, and, potentially, rely on the community as well as crowdsourcing for manual validation.

Because the MASC is an open resource that the community can continually enhance with additional annotations and modifications, the project should serve as a model for community-wide resource development. Past experience with corpora such as the *Wall Street Journal* shows that the community is eager to annotate available language data; MASC, which includes language data covering a range of contemporary genres, should provide an even more appealing base for a global community- and contribution-driven annotation effort. We share the vision of the LLOD cloud to create a massive, inter-linked linguistic infrastructure for the study and processing of human languages, for example, by linking MASC's WordNet and FrameNet annotations to those resources as well as wordnets and framenets in other languages as well as resources such as BabelNet [25], thus creating a global resource for multi-lingual technologies. MASC is intended to serve as a step in achieving that vision.

References

1. Baker, C.F., Fellbaum, C.: WordNet and FrameNet as complementary resources for annotation. In: Proceedings of the Third Linguistic Annotation Workshop, pp. 125–129. Association for Computational Linguistics, Suntec, Singapore (2009). URL http://www.aclweb.org/anthology/W/W09/W09-3021
2. Blumtritt, J., Bouda, P., Rau, F.: Poio API and GraF-XML: A radical stand-off approach in language documentation and language typology. In: Proceedings of Balisage: The Markup Conference 2013, *Balisage Series on Markup Technologies*, vol. 10. Montreal, Canada (2013). DOI 10.4242/BalisageVol10.Bouda01
3. Chiarcos, C., Hellmann, S., Nordhoff, S.: Linking linguistic resources: Examples from the open linguistics working group. In: C. Chiarcos, S. Nordhoff, S. Hellmann (eds.) Linked Data in Linguistics, pp. 201–216. Springer (2012)
4. Chiarcos, C., Ritz, J., Stede, M.: By all these lovely tokens... merging conflicting tokenizations. Language Resources and Evaluation **46**(1), 53–74 (2012)
5. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of ACL'02 (2002)
6. Fellbaum, C., Baker, C.: Aligning verbs in WordNet and FrameNet. Linguistics (to appear)
7. 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). DOI http://dx.doi.org/10.1017/S1351324904003523
8. Fillmore, C.J., Jurafsky, D., Ide, N., Macleod, C.: An american national corpus: A proposal. In: Proposal. Proceedings of the First Annual Conference on Language Resources and Evaluation. Paris: European Language Resources Association, pp. 965–969 (1998)

9. Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N.: Offspring from reproduction problems: What replication failure teaches us. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1691–1701. Association for Computational Linguistics, Sofia, Bulgaria (2013)
10. Ide, N.: An open linguistic infrastructure for annotated corpora. In: I. Gurevych, J. Kim (eds.) he Peoples Web Meets NLP: Collaboratively Constructed Language Resources, pp. 263–84. Springer (2013)
11. Ide, N., Bonhomme, P., Romary, L.: Xces: An xml-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association (2000)
12. Ide, N., Reppen, R., Suderman, K.: The american national corpus: More than the web can provide. In: Proceedings of the Third Language Resources and Evaluation Conference, Las Palmas, pp. 839–844 (2002)
13. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Natural Language Engineering **10**(3-4), 211–225 (2004). DOI http://dx.doi.org/10.1017/S135132490400350X
14. Ide, N., Romary, L.: Representing linguistic corpora and their annotations. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC 2006 (2006)
15. Ide, N., Suderman, K.: Integrating linguistic resources: The american national corpus model. In: Proceedings of the Fifth Language Resources and Evaluation Conference (LREC). Genoa, Italy (2006)
16. Ide, N., Suderman, K.: GrAF: A graph-based format for linguistic annotations. In: Proceedings of the Linguistic Annotation Workshop, pp. 1–8. Association for Computational Linguistics, Prague, Czech Republic (2007). URL http://www.aclweb.org/anthology/W/W07/W07-1501
17. Ide, N., Suderman, K.: The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. Language Resources and Evaluation (2014)
18. Ide, N., Suderman, K., Simms, B.: ANC2Go: A web application for customized corpus creation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC). European Language Resources Association, Valletta, Malta (2010)
19. ISO: Language Resource Management - Linguistic Annotation Framework. iso 24612 (2012)
20. Kilgarriff, A.: Googleology is bad science. Computational Linguistics **33**(1) (2007)
21. Kremer, G., Erk, K., Pad, S., Thater, S.: What substitutes tell us – analysis of an "all-words" lexical substitution corpus. In: Proceedings of the Conference of the European Association for Computational Linguistics. Gothenburg, Sweden (2014)
22. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Description and construction of text structures. In: G. Kempen (ed.) Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics, pp. 85–95. Nijhoff, Dordrecht (1987)
23. 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)
24. Moro, A., Navigli, R., Tucci, F.M., Passonneau, R.J.: Annotating the masc corpus with babelnet. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
25. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
26. 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). DOI http://dx.doi.org/10.1109/ICSC.2007.67