Corpus Development and Publication

Andrew W. Cole
University of Pennsylvania
Linguistic Data Consortium
Suite 810
3600 Market Street
Philadelphia Pennsylvania 19104
andrew.cole@ldc.upenn.edu

Abstract

This paper will discuss issues relevant to corpus development and publication at the LDC and will illustrate those issues by examining the history of three LDC corpora. This paper will also briefly examine alternative corpus creation and distribution methods and their challenges. The intent of this paper is to increase the available linguistic resources by describing the regulatory and technical environment and thus improving the understanding and interaction between corpus providers and distributors.

1. Introduction

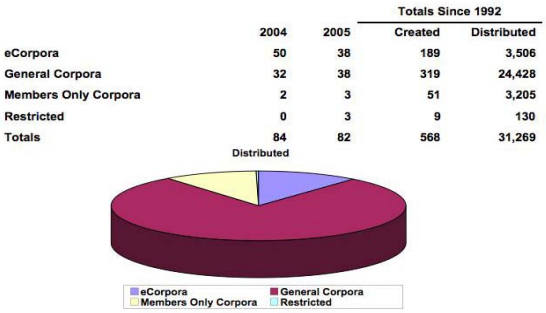
The Linguistic Data Consortium (LDC) is an open consortium of universities, companies, and government research laboratories that creates and distributes speech and text databases, lexicons, and other resources. The University of Pennsylvania is the host institution for the LDC. The LDC was founded in 1992 with a grant from the Defense Advanced Research Projects Agency (DARPA). Currently, all LDC publication and distribution activities are self-supporting, while new data creation is typically supported by grants for that purpose.

Each year the Linguistic Data Consortium (LDC) receives an average of fifty corpus submissions from external researchers. In addition, the LDC creates an additional fifty corpora that are released initially for use within specific research communities before they are released generally.

From this pool, the LDC publishes an average of three General Release corpora per month. These General Release corpora are available to our Members and, where allowed by intellectual property rights (IPR), to non-member licensees. The LDC also releases a highly variable number of eCorpora to research groups each year. eCorpora, due to rapid collection and distribution requirements, do not receive the same level of documentation review and quality control of data as General Release corpora.

This paper will describe the regulatory and technical environment under which the LDC receives and processes corpora with the intent to increase the number and quality of linguistic resources by improving the interaction between corpus providers and distributors. Finally this paper will explore alternatives to centralized agency distribution and list some of the opportunities and challenges in this area.

To provide a rough idea of the individual corpora and number of copies released by the LDC, Figure 1, lists these counts for the last two years as well as the cumulative number of corpora released and copies distributed since the LDC's inception in 1992.



	2004	2005	Totals Since 1992 Created	Distributed
eCorpora	50	38	189	3,506
General Corpora	32	38	319	24,428
Members Only Corpora	2	3	51	3,205
Restricted	0	3	9	130
Totals	84	82	568	31,269

Figure 1

2. Description

In practice, corpora consist of two initial components, data and metadata. Data can be (but are not limited to) the following categories:

1. Text
2. Audio
3. Video
4. Lexicons

Here I use the term metadata in the straightforward sense of information about data Corpus metadata can be (but are not limited to) the following:

1. Unique Name (obvious but non-trivial in practice)
2. Description
3. Project or Sponsors
4. Primary Contact
5. Author(s)
6. Data type (per above paragraph)
7. Data Source Providers
8. Size
9. Encoding
10. Suggested Uses
11. Applicable Standards
12. Quality Control
13. Collection/Annotation Specifications

Many of the early corpora released through the LDC lacked some of the metadata listed above. One possible cause was the close connection between the producers and probable consumers of the data such that much of this metadata was intimately known to both the corpus developers and the LDC staff involved. Over time, however, as the range of researchers with often unanticipated uses for LDC corpora expanded, the need for more metadata was empirically determined – by receipt of numerous requests for the meta-data listed above. Another cause is the prevailing attitude toward corpora. They are still seen among some communities as fundamentally different from other scholarly products so that the idea of assigning authorship, for example, remains foreign to some.

3. Corpora Relationships

Over time, as the range and number of LDC corpora expanded, a hierarchical relationship developed within the corpora. For example, a telephone speech (audio) corpus would be released, the LDC or other researchers would transcribe the audio and provide a turn segmented transcript, another researcher would further segment the transcript by sentences and still yet another researcher would annotate the segmented transcript for disfluencies or entities and relations, all leading to a tree structure as illustrated below:

telephone speech (audio)
- turn aligned transcription
  - time aligned transcription
  - sentence aligned transcription
    - disfluencies
    - entity
      - entity relation
    - … *ad infinitum* …

The possibility of corrections to any of the annotations or to the source audio could result in the possibility of a cycle in the relationship graph with the attendant problem of determining which source applied to which annotation. This problem was administratively eliminated by assigning versions to corrections and to any new releases of similar or related corpora, thus:

telephone speech (audio)
- turn aligned transcription
  - time aligned transcription
  - sentence aligned transcription
    - disfluencies
    - entity
      - entity relation
- **turn aligned transcription version x.y**
  - **time aligned transcription of version x.y**

Where these annotations are embedded via `<sgml>` tags within the source text itself, a further complication arises if a researcher wished to view more than one annotation at a time, such as both disfluencies and entity relationships. Stand-off annotation ameliorates this problem to some extent as does a consistent model for structuring annotation. LDC has promoted structured models of annotation in the form of Annotation Graphs (Bird, Liberman 2001) in which annotations are written to separate files and linked by arc offsets to the source data.

In summary, to make a somewhat labored comparison, corpora have both atomic elements; data and metadata, which can be combined to build extended molecular structures based on particular source corpora.

4. Standards

The use of standard systems of encoding and annotation initially provided the LDC's core research communities with data that was well understood and immediately useful for research. However, as the range of uses of LDC data increased and as other research communities provided data to be released through the LDC, a wider variety of standards became applicable and the selection of which standard to apply to what data and how intensively to apply the applicable standards became something of a Cartesian headache for the LDC. Over time the LDC has come to rely on the interested research communities to provide annotations using alternative standards or in greater depth than the LDC is financially able to create.

In general, which standard to follow and how deeply to adhere to that standard, are complex empirical problems which are affected by financial and research requirements. For example, IMDI, (Wittenburg, Broeder, Sloman 2000) represents a rich metadata specification that has been proposed as a general standard. Time and fiscal constraints as well as the traditions of different research communities make it impossible to adopt a single standard for all corpora.

As an example, although researchers may desire to provide IMDI style metadata for many of its corpora, time, funds and their practice may force them into a Hobson's choice of releasing a corpus conforming to part of IMDI or not being able to release it at all. This is truly a case of, as Voltaire said, "The best is the enemy of the good". The best standards, applied rigorously, would delay, and possibly bar, the release of data.

5. Intellectual Property Rights (IPR) and Informed Consent

Perhaps no area of data collection and distribution is currently as unstable as the acquisition of rights to release data. Research in the United States involving human subjects must be approved by a institutional review board (IRB) that is itself federally approved. Commercial producers of text, audio, and video data are naturally sensitive to any release of their data that would adversely affect their competitive position and will generally only release data under restrictive licenses. Finally, researchers themselves often wish to maintain some control over data that they release, if only to ensure that the data continues to be available without subsequent restriction.

Conversely, libraries, researchers, and even Google™ feel that small amounts of unrestricted data use should be allowed under such legal theories as fair use.

The LDC, as a central data repository with attendant responsibilities, receives data collected under various IRB protocols and negotiates IPR agreements with numerous data providers and thus acts as an intermediary for agreements. However, the LDC can not, and does not, act

as a legal representative for either researchers or data providers.

In practice when we acquire data from providers for research purposes we only provide that data to researchers under the same restrictions as we receive it.

6. Quality Assurance

Quality assurance of linguistic resources is closely related to standards. The goal of quality assurance is two-fold, to provide usable data and to provide researchers with the means to replicate the quality assurance process.

In practice, the ability to replicate automated quality control procedures implies that the data is usable as researchers can run quality assurance processes against the data when received to ensure that there are no media or transmission errors.

For each type of data a different method of quality control inheres. The simplest, text data, are generally validated via programs such as NSGMLS (http://www.jclark.com/sp/nsgmls.htm) via an associated Document Type Definition (DTD). However, the LDC is moving towards using XML and associated Schema, which can be validated via tools such as Saxon (http://sourceforge.net/projects/saxon). For more complex annotation systems, applicable quality assurance processes are requested from the data provider. Unfortunately, in some cases, no such validation tool or control document exists and the data is released with suitable notice to that effect.

For example, manual (human) annotation relying on judgment, generally use multiple pass with adjudication to improve corpus quality.

For audio data, quality control generally relates to signal quality and correcting or documenting issues such as clipping, sample size, sample frequency, and header encoding. In the past, most LDC audio data was released in sphere format, however, we are now considering the possibility of releasing a larger percentage of our audio data in RIFF (wav) format. Metadata quality control for audio is more problematic. For example, ensuring that a speaker is actually from the southern United States relies on the speaker's veracity.

For video data, the standards are still somewhat fluid and have not yet settled, leaving the decision of video encoding to the researcher providing the data or the sponsor funding the collection.

7. Processing and Distribution

In general, however, most corpora come from researchers or are created in response to research needs and are processed by the LDC and distributed to LDC Members, researchers, or project participants.

After the corpora are received by the LDC External Relations group, both the data and documentation is reviewed for conformance to standards and quality control as described in the documentation. Further, IPR is reviewed and re-verified, as initial corpus design is often altered during collection and development which result in changes in source or epoch that mandate a final confirmation.

Finally the corpus is prepared for distribution via HTTP file download or on media such as CD, DVD, or hard drive. Random samples of the media are examined for duplication errors.

An announcement is then made to the relevant communities and the corpus is shipped after appropriate agreements and payment are received.

8. Example Corpora

8.1. Arabic Gigaword, Second Edition

8.1.1. Description

Arabic Gigaword, Second Edition, LDC Catalog number LDC2006T02 (Graff, et al, 2006) was created to provide substantial Arabic newswire resources to the research community. LDC Staff David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda produced this second edition.

The Second Edition contains over 1.5 billion words from Agence France Presse, Al Hayat,An Nahar, Xinhua and Ummah Press.

Figure 2 shows an example document from the Second Edition corpus.

Figure 2

8.1.2. Related Corpora

The initial Arabic Gigaword corpus, LDC2003T12 (Graff, 2003) was released in 2003 with 390 million words of Arabic newswire from the following sources: Agence France Presse, Al Hayat, An Nahar, Xinhua.

Additional versions of Arabic Gigaword corpora will be released as significant additional data become available.

8.1.3. Standards

Although each newswire source encoded the data in different formats with different document tagging, all data was standardized to UTF-8 encoding with the anticipation that this would allow the data to be used by a wide variety of researchers. Also, a reference DTD is provided to validate the document tagging.

In addition, file naming was standardized and conforms to structure and source grouping in other large text corpora such as the Chinese and English Gigaword corpora.

8.1.4. Intellectual Property Rights (IPR)

The IPR for each of these sources was obtained by the LDC through long term agreements for research use and distribution of the newswire. These agreements allowed the LDC, for a fee, to redistribute this data to Member and non-Member researchers (with the completion of the appropriate agreement between the researcher and the LDC).

In addition, the Trustees of the University of Pennsylvania retained copyright for annotations and processing by LDC Staff.

8.1.5. Quality Assurance

Quality assurance involved processing all files with NSGML to conform to an included DTD to ensure that all files contained the predicted encoding and that all tagging conformed to the rules embodied in the DTD.

8.1.6. Processing and Distribution

To date 39 copies have been distributed to LDC Members and associated researchers via a single DVD. The data is compressed via GNU GZip to 1.4 Gbytes.

8.2. ACE 2005 Multilingual Training Corpus

8.2.1. Description

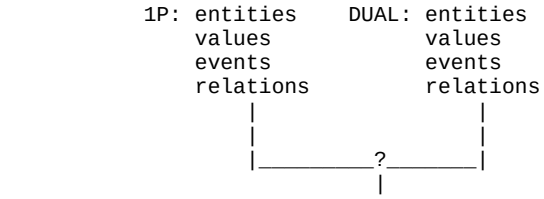
ACE 2005 Multilingual Training Corpus, LDC Catalog number LDC2006T06 (Walker, et al, 2006) was developed to provide English, Arabic and Chinese training data for the 2005 Automatic Content Extraction (ACE) technology evaluation. The corpus consists of data of various types, annotated for entities, relations and events and was previously distributed as an eCorpus (LDC2005E18) to participants in the 2005 ACE evaluation.

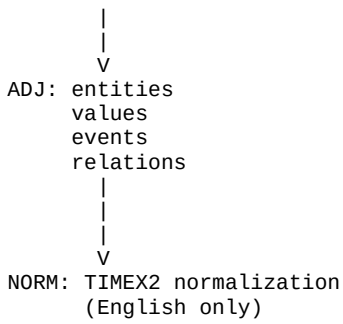
LDC Staff Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda prepared the data and assembled the annotations from the Agence France Presse, The Associated Press, New York Times, Xinhua News Agency, Cable News Network LP, LLLP, SPH AsiaOne Ltd, China Broadcasting System,China National Radio, China Television System, China Central TV, Al Hayat, An-Nahar, Nile TV.

The corpus contained several different genres, from Newswire to Broadcast News in three languages with the following counts:

English	303,833 words
Chinese	334,121 characters
Arabic	112,233 words

The full annotation process for 2005 is represented below:





An example of a partial annotation of an English file is presented below:

```
<?xml version="1.0"?>
<!DOCTYPE source_file SYSTEM "apf.v5.1.1.dtd">
<source_file URI="CNN_ENG_20030630_085848.18.sgm"
SOURCE="broadcast news" TYPE="text" AUTHOR="LDC"
ENCODING="UTF-8">
<document DOCID="CNN_ENG_20030630_085848.18">
<entity ID="CNN_ENG_20030630_085848.18-E1" TYPE="GPE"
SUBTYPE="State-or-Province" CLASS="SPC">
  <entity_mention ID="CNN_ENG_20030630_085848.18-E1-1"
TYPE="NAM" LDCTYPE="NAM" ROLE="LOC">
    <extent>
      <charseq START="82" END="91">california</charseq>
    </extent>
    <head>
      <charseq START="82" END="91">california</charseq>
    </head>
  </entity_mention>
```

8.2.2. Related Corpora

During the ACE evaluation this corpus was distributed as eCorpus LDC2005E18, to participants in the 2005 ACE evaluation.

8.2.3. Standards

Due to the manual nature of the annotations, a rigorous and complex quality control process was implemented such that training data files for all languages are dually annotated for all tasks by two annotators working independently.

The first pass (complete) annotation is called 1P and the independent dual first pass (complete) annotation is called DUAL. For both 1P and DUAL, a single annotator completes all tasks (entities, values, relations & events) for a file. Files are assigned via an automated Annotation Workflow System (AWS), and file assignment is double-blind.

Discrepancies between the 1P and DUAL version of each file are then adjudicated by a senior annotator or team leader, resulting in a high-quality gold standard file. The gold standard adjudicated file is known as ADJ. After adjudication, TIMEX2 values are normalized for English only.

8.2.4. Intellectual Property Rights (IPR)

Similar to the Gigaword corpus, the IPR for each of these sources was obtained by the LDC through long term agreements for research use and distribution of the newswire and allowed the LDC, for a fee, to redistribute this data to Member and non-Member researchers (with the completion of the appropriate agreement between the researcher and the LDC).

In addition, the Trustees of the University of Pennsylvania retained copyright for annotations by LDC Staff.

8.2.5. Quality Assurance

Listed below is a subset of the extensive quality control procedures and structural features used to perform quality control on this corpus.

- All non-complex entity mentions have heads. For APF, this means that all entity mentions have heads
- No English passages are annotated in non-English files
- All relation mentions have exactly two non-timex2 arguments
- All relation arguments are contained in the extent of the relation mention
- All event mention arguments are contained in the extent of the event mention
- All NAMPRE and NOMPRE GPE mentions have GPE as their role
- No relations have mentions from the same entity as their only non-timex2 arguments
- All files have exactly one timex2 annotation in the DATETIME field
- No annotation extents overlap without nesting (entity mention, relation mention, event mention, value mention, entity mention head, event mention anchor)
- There are no annotations inside of sgm tags
- There are no instances where an entity and an event share exactly the same head/anchor
- All relation arguments have types that are allowed for their argument position based on their entity/value type
- All event arguments have types that are allowed for their role based on their entity/value type
- All entities, values, relations, events have permissible type-subtype pairs
- All files successfully convert to APF
- All APF files validate against DTD
- All APF files can be scored against themselves
- All instances of cross-type metonymy manually reviewed
- All instances of co-extensive entity mentions with the same heads manually reviewed
- Check for event mentions whose anchor is the full extent of the mention
- Manual scan of all PRO extents for outliers in adjudicated files
- Manual scan of all NOM heads with different entity type/subtype values in different parts of the corpus (adjudicated files only)
- Manual scan of all NAM heads with different entity type/subtype values in different parts of the corpus (adjudicated files only)
- Manual scan of all relation mentions by relation type/subtype and argument type/subtype for outliers in adjudicated files
- Manual and automatic scans of mention extents by patterns to identify inconsistencies in adjudicated files
- Search for relations with frequently-confused types based on argument types (in particular, PHYS.Located vs. ART.UOIM and ORG-AFF.Employment vs. GPE-AFF.CRRE)
- Search for co-extensive, co-related event mentions
- Scan all clitic pronoun mentions that are not participants in the event whose anchor they are attached to (Arabic)
- Scan all unannotated common TIMEX2 triggers (English)
- Manually examine and correct or describe all fatal errors and warnings generated by the most recent version of the scorer

8.2.6. Processing and Distribution

To date, 31 copies have been distributed to LDC Members and associated researchers. This corpus was released on a single DVD.

8.3. Buckwalter Arabic Morphological Analyzer Version 2.0

8.3.1. Description

The Buckwalter Arabic Morphological Analyzer Version 2.0, LDC Catalog number LDC2004L02 (Buckwalter, 2004) is intended to support researchers needing an Arabic lexicon and morphological analysis. Tim Buckwalter produced this corpus at the LDC.

Version 2.0, was released in 2004 and contained 78,839 entries representing 40,219 lemmas as well as AbuMorph, an enhanced tool for morphology and Part of Speech (POS) tagging with example input text and output XML files, as follows:

1. Three lexicon files: dictPrefixes, dictStems, and dictSuffixes.
2. Three compatibility tables: tableAB, tableAC, and tableBC.
3. Perl code (AraMorph.pl) that makes use of the three lexicon files and three compatibility tables in order to perform morphological analysis and POS-tagging of Arabic words.
4. Sample Arabic input file (infile.txt) in Windows 1256 encoding.
5. Sample morphology analysis output file (outfile.xml) in Unicode UTF-8 encoding.
6. Documentation in readme.txt

8.3.2. Related Corpora

The Buckwalter Arabic Morphological Analyzer Version 2.0 used the following corpora for validation:

Arabic Treebank: Part 1 version 2.0 (LDC2003T06) (Maamouri et al, 2003)
Arabic Treebank: Part 2 version 2.0 (LDC2004T02) (Maamouri et al, 2004)
Arabic Treebank: Part 3 version 1.0 (LDC2004T11) (Maamouri et al, 2004)

The Buckwalter Arabic Morphological Analyzer Version 1.0, LDC Catalog number LDC2002L49 (Buckwalter, 2002) was released in 2002 and contained 82,158 entries representing 38,600 lemmas.

8.3.3. Standards

The encoding listed below is an effort to provide a method for automatic morphological analysis and is composed of entries in the three lexicon files which consist of the following four tab-delimited fields:

1. the entry (prefix, stem, or suffix) WITHOUT short vowels and diacritics
2. the entry (prefix, stem, or suffix) WITH short vowels and diacritics
3. its morphological category (for controlling the compatibility of prefixes, stems, and suffixes)
4. its English gloss(es), including selective POS data within XML tags <pos>...</pos>

XML tagging and encoding in UTF-8 was used where applicable for the data and the AbuMorph system generates XML data in UTF-8.

8.3.4. Intellectual Property Rights (IPR)

This corpus was produced at the LDC and the Trustees of the University of Pennsylvania retain copyright for research distribution. This corpus is only available to LDC Members.

Commercial rights to this corpus may be obtained through:

QAMUS LLC
448 South 48th St.
Philadelphia, PA 19143
ATTN: Tim Buckwalter
email: license@qamus.org

8.3.5. Quality Assurance

The Buckwalter Morphological Analyzer registered coverage rates as follows:

90% accuracy in the analysis of the Arabic Treebank: Part 1 version 2.0 (LDC2003T06),

99.24% accuracy in the analysis of the Arabic Treebank: Part 2 version 2.0 (LDC2004T02)

99.25% accuracy in the analysis of the Arabic Treebank: Part 3 version 1.0 (LDC2004T11)

8.3.6. Processing and Distribution

To date 27 copies have been distributed via FTP and HTTP download. This corpus is only available to LDC Members.

9. Alternative Creation and Distribution

Although the LDC is comfortable with current corpus development and distribution methods, with the advent of the internet and developing web annotation tools, there is the opportunity to extend corpus development and distribution beyond centralized repositories.

The opportunity for distributed annotation of linguistic resources via web browser obviously offer advantages of assembling geographically separate researchers to work together on common source data.

Further there is the real possibility (perhaps already reality) of training interested non-specialists to provide source data and annotations. For example http://www.librivox.org/ solicits volunteers who have created over twenty-five English language recordings of books in the public domain and at least five similar German language recordings.

This model is adaptable for research collection akin to the model the LDC now uses to locate and record telephone speakers.

Distribution, IPR and quality control issues will still need to be addressed for internet based transcriptions, recordings and annotations. For example, some entity, similar to librivox, will need to identify source materials to be recorded or annotated and verify that IPR is obtained for any source not in the public domain as well as ensuring that all donated work is covered by suitable IPR agreements.

In addition, some level of quality control or validation will need to be established and some auditing method implemented in order to ensure that resources obtained are usable by researchers. Implementing a process similar to that described for the ACE 2005 Multilingual Training Corpus would require some method to compare multiple annotation streams to render a "standard" annotation.

These challenges, however, seem well within the range of possibility and the LDC will pursue them as opportunities present themselves in current and future data collection and annotation projects.

10. References

Steven Bird and Mark Liberman (2001). *A formal framework for linguistic annotation* Speech Communication 33(1,2), pp 23-60.

Wittenburg, P., Broeder, D., and Sloman, B., (2000), *EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper.* LREC 2000 Workshop, Athens.

Graff, D., Chen, K., Kong, J., and Maeda, K., (2006), *Arabic Gigaword Second Edition.* Linguistic Data Consortium, Philadelphia.

Graff, D., (2003), *Arabic Gigaword* Linguistic Data Consortium, Philadelphia.

Walker, C., Strassel, S., Medero J., and Maeda, K., (2006) *ACE 2005 Multilingual Training Corpus.* Linguistic Data Consortium, Philadelphia.

Buckwalter, T, (2004) *Buckwalter Arabic Morphological Analyzer Version 2.0.* Linguistic Data Consortium, Philadelphia.

Maamouri, M., Bies, A., Jin, H., Buckwalter, T, (2003) *Arabic Treebank: Part 1 v 2.0* Linguistic Data Consortium, Philadelphia.

Maamouri, M., Bies, A., Buckwalter, T, Jin, H., (2004) *Arabic Treebank: Part 2 v 2.0* Linguistic Data Consortium, Philadelphia.

Maamouri, M., Bies, A., Buckwalter, T, Jin, H., (2004) *Arabic Treebank: Part 3 v 1.0* Linguistic Data Consortium, Philadelphia.

Buckwalter, T, (2002) *Buckwalter Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium, Philadelphia.