LDC: ANNOTATION Treebanks

Introduction

Treebanks are widely used in the Natural Language Processing (NLP) community to support the creation and training of parsers and taggers, work on machine translation and speech recognition, and research on joint syntactic and semantic role labeling. Treebanks have also been used as the basis for downstream annotation projects, such as PropBanks, the Penn Discourse Treebank and word alignment.

What is a Treebank?

Treebanks are fully parsed corpora that are manually annotated for syntactic structure at the sentence level and for part-of-speech or morphological information at the token level. Every token and every sentence in the text is annotated.

What is the Goal of Treebanking?

To represent useful linguistic structure in an accessible way, including:

- Consistent annotation
- Searchable trees
- "Correct" linguistic analysis if possible, but at a minimum consistent and searchable
- Annotation useful to both linguistic and NLP communities
- Empirical methods providing portability to new languages
- Structures that can be used as the basis for additional downstream annotation

The association of treebanks with the Linguistic Data Consortium (LDC) began with the publication of the original Penn English Treebank in 1995. LDC continues treebanking efforts today with the publication and in house annotation and publication of both Arabic and English Treebanks.

LDC takes advantage of a pragmatic mixture of manual, semi-automatic and fully automatic annotation methods. In addition, LDC uses and continues to support the development of robust tools such as morphological analyzers and parsers to speed up the annotation process.

LDC's Catalog contains many treebank publications produced by NLP researchers. These include the Prague Dependency Treebanks, Chinese Treebank, Korean Treebank and OntoNotes.

```
( S  ( CONJ  و  wa)
     ( VP  ( PRT  ( FUT_PART  س  sa ))
           ( IV3MS+IV+IVSUFF_MOOD:I  يرافق  yu+rAfiq+u  )
           ( NP- OBJ  ( NOUN_PROP  باول  bAwil ))
           ( NP- SBJ  ( NOUN+CASE_INDEF_NOM  وفد  wafod+N )
                      ( ADJ+CASE_INDEF_NOM  اعلامي  <iEolAmiy~+N  )
                      ( ADJ+C ASE_INDEF_NOM  اميركي  >amiyrokiy~+N  )))
     ( PUNC  .))
```

```
( S  ( NP- SBJ  ( DT  An  )
                ( JJ  American  )
                ( NN  media  )
                ( NN  delegation  ))
     ( VP  ( MD  will  )
           ( VP  ( VB  accompany  )
                 ( NP  ( NNP  Powell  ))))
     ( .  .))
```

Parallel Arabic and English Treebanked Sentence

Arabic Treebank at LDC

The Penn Arabic Treebank (ATB) project began in 2001 at LDC with the initial support of the DARPA TIDES program and later of the DARPA GALE program. ATB corpora are annotated for morphological information, part-of-speech and English gloss, all at the token level, and for syntactic structure in the Penn Treebank 2 style. Current ATB activities include a team of Arabic annotators in Tunisia.

In addition to the expected issues associated with the complex annotation of data, LDC encountered a number of issues that are specific to a highly inflected language with a rich history of traditional grammar. In order to design an annotation system for Arabic, LDC relied on traditional Arabic grammar, previous grammatical theories of Modern Standard Arabic and modern approaches, and especially the Penn Treebank approach to syntactic annotation. LDC was innovative with respect to traditional grammar when necessary and when other syntactic approaches were found to account for the data.

The LDC Arabic Treebank team has significantly revised and enhanced its annotation guidelines and annotation procedures during GALE. LDC is currently publishing both revised versions of existing ATB corpora and new corpora annotated according to the revised guidelines.

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

```
INPUT_STRING: جنديأ
    IS_TRANS: jndyAF
       INDEX: P1W2
     OFFSETS: 4-11
      TOKENS: P1W2-P1W2
      STATUS: 1
       LEMMA: [junodiy~_1]
 UNSPLITVOC: (junodiy~AF)
         POS: NOUN+CASE_INDEF_ACC
         VOC: junodiy~+AF
       GLOSS: soldier + [acc.indef.]
```

Arabic Treebank Token Level Annotation

Improved ATB guidelines, improving inter-annotator agreement and continuing improvement in parsing scores are the result of a fruitful collaboration between data producers, sponsors and end users. It is expected that continued collaboration will benefit both annotation production and future NLP applications. Current Arabic Treebank Guidelines can be found at projects.ldc.upenn.edu/ArabicTreebank/

English Treebank at LDC

LDC's English treebank focus has evolved over time. Initially, LDC published the Penn English Treebanks (2 and 3) and continues to do so today. The need for new varieties of English treebank data has continued to grow, and LDC expanded its expertise to address new treebank research.

LDC's current English treebank research focuses on parallel data translated from Arabic, which is also annotated by the Arabic Treebank team, and includes broadcast news, broadcast conversation, web text and dialect data. The English input files (translated from the Arabic) are first automatically part-of-speech tagged and parsed. The tags and parses are then hand corrected by English Treebank annotators.

This annotation pipeline is followed by a quality control process consisting of a series of specific searches for potential inconsistency and parser or annotation error. Currently there are approximately 100 types of inconsistency searches and this number will continue to increase. Any errors found in these searches are hand corrected.

Recent English Treebank publications include two corpora of biomedical texts, parallel data translated from the Chinese Treebank and a corpus of conversational telephone speech that is also annotated for structural metadata (MDE).

All English (Translation) Treebank data has recently been revised and updated according to GALE community request. The guidelines for part-of-speech and syntactic annotation are essentially Penn Treebank 2 style, updated to comply with annotation guidelines agreed upon by GALE sites (including tokenization changes and the "Treebank-PropBank merge" or "Treebank IIa" changes). Revised English Translation Treebank corpora will be made available as LDC publications.

English Treebank Guidelines can be found at: projects.ldc.upenn.edu/gale/task_specifications/EnglishXBank/

Arabic and English Treebank Data Released by LDC

LDC has published versions of over 25 treebank corpora in five languages: Arabic, English, Chinese, Czech and Korean. Approximately half of these corpora, primarily Arabic and English data, were annotated in house while the remainder were created by third party authors. The chart below lists the current totals of Arabic and English treebank data LDC has released.

	Arabic	English	Parallel Data
Newswire (NW)	750 K tokens	2 million tokens	500 K tokens
Broadcast News (BN)	400 K tokens	400 K tokens	400 K tokens
Broadcast Conversation (BC)	Planned for 2010	Planned for 2010	Planned for 2010
Conversational Telephone Speech (CTS)	45 K tokens	500 K tokens	Pending
Web Text (WB)	100 K tokens	100 K tokens	100 K tokens
Other	Undetermined	1 million tokens	Undetermined
Totals	1.2 million tokens	4 million tokens	1 million tokens

Future Plans

Treebank annotation is broadening to include Arabic dialect and Arabic speech corpora along with parallel English translation data to support the development of dialect parsers, speech recognition and machine translation efforts. Parallel Arabic and English Treebank speech corpora (broadcast news and broadcast conversation) and web data will soon be available in the LDC Catalog.

For more infomation on available publications and news on upcoming releases, please contact LDC's Membership Office at ldc@ldc.upenn.edu