Large Multi-lingual, Multi-level and Multi-genre Annotation Corpus

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Summerville Conger, Stephen Grimes, Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania
Philadelphia, PA 19104 USA
Department of Linguistics and Computer Science, University of Colorado
Boulder, CO 80309 USA
Computer Science Department, Brandeis University
Waltham, MA 02453 USA
Raytheon BBN Technologies
Cambridge, MA 02138 USA
Email:{xuansong,maamouri,bies,sgrimes,strassel}@ldc.upenn.edu{martha.palmer,kathryn.conger@colorado.edu
xuen@brandeis.edu lramshaw@bbn.com

Abstract

High accuracy for automated translation and information retrieval calls for linguistic annotations at various language levels. The plethora of informal internet content sparked the demand for porting state-of-art natural language processing (NLP) applications to new social media as well as diverse language adaptation. Effort launched by the BOLT (Broad Operational Language Translation) program at DARPA (Defense Advanced Research Projects Agency) successfully addressed the internet information with enhanced NLP systems. BOLT aims for automated translation and linguistic analysis for informal genres of text and speech in online and in-person communication. As a part of this program, the Linguistic Data Consortium (LDC) developed valuable linguistic resources in support of the training and evaluation of such new technologies. This paper focuses on methodologies, infrastructure, and procedure for developing linguistic annotation at various language levels, including Treebank (TB), word alignment (WA), PropBank (PB), and co-reference (CoRef). Inspired by the OntoNotes approach with adaptations to the tasks to reflect the goals and scope of the BOLT project, this effort has introduced more annotation types of informal and free-style genres in English, Chinese and Egyptian Arabic. The corpus produced is by far the largest multi-lingual, multi-level and multi-genre annotation corpus of informal text and speech.

Keywords: machine translation; parallel aligned Treebank; word alignment; PropBank; co-reference

1. Introduction

Annotated corpora constitute a crucial resource for developing NLP systems and facilitating linguistic analysis of languages. With the rapid growth of the internet, NLP technologies are challenged with an avalanche of unstructured user-generated data. Off-the-shelf tools, trained on venerable formal data, perform worse when applied to new social media data. Various research and development efforts have been invested in this new domain -- massive informal and unstructured data. In this line, Ryan Cotterell and Chris Callison-Burch (2014) created a multi-dialect and multi-genre corpus of informal text via Amazon's Mechanical Turk services. With a crowdsourcing approach, Jha et al. (2010) built a prepositional phrase attachment corpus of informal and noisy blog text. Owoputi et al. (2013) created part-of-speech tagged data for informal and online conversational Twitter text. The OntoNotes corpus (Weischedel et al. 2013) is a collaborative effort between BBN Technologies, Brandeis University, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences. The OntoNotes corpus comprises integrated annotation of multiple levels in various genres and in three languages (English, Chinese, and MSA Arabic), providing structural information (syntax and predicate argument structure) and shallow semantics (word sense and coreference).

ANNOTATION DATA VOLUME (Words)				
DF	**WA**	**TB**	**PB**	**CoRef**
Chinese	481,024	421,121	419,623	421,121
Egyptian	400,448	400,448	400,448	400,448
English	881,472	268,907	268,907	268,907
EngTrans	147,433	147,433	147,433	147,433
SMS/Chat	**WA**	**TB**	**PB**	**CoRef**
Chinese	388,027	303,640	303,640	149,246
Egyptian	349,414	349,414	198,007	102,864
English	737,441	117,054	117,054	117,054
EngTrans	212,620	212,620	159,860	n/a
CTS	**WA**	**TB**	**PB**	**CoRef**
Chinese	160,153	160,153	120,897	120,897
Egyptian	153,171	153,171	99,201	99,201
English	313,324	n/a	n/a	n/a
EngTrans	170,526	175,133	109,816	109,816
Total	**WA**	**TB**	**PB**	**CoRef**
Chinese	1,029,204	884,914	844,160	691,264
Egyptian	903,033	903,033	697,656	602,513
English	1,932,237	385,961	385,961	385,961
EngTrans	530,579	535,186	417,109	257,249

Table 1: BOLT Annotation Data

To leverage overall NLP technologies to a new horizon, DARPA initiated the BOLT program, encouraging development and adaptation of systems for automated translation and analysis for informal genres of text and speech in online and in-person communication. As a part of the BOLT program, LDC developed large volume fine-grained annotation resources comprised of four types of annotation in three languages and in three informal genres, as indicated in Table 1 (DF stands for discussion forum, SMS/chat for text message and chat, CTS for conversational telephone speech, and EngTrans for English Translation). Both “EngTrans for WA” (indicating parallel-aligned Treebank annotation) and “English for WA” are based on the source word count. Up to date, this is the largest annotation corpus of informal text and speech.

This paper presents the process for creating the corpus, highlighting approaches and challenges with informal data in different languages. Section 1 is the introduction. Section 2 delineates the annotation pipeline. Section 3 focuses on methodologies for each type of annotation and data structure. Section 4 describes data consistency and integrity across different types of annotation. Section 5 concludes the paper.

2. Annotation Pipeline

2.1 Source Data for Treebank Annotation

Annotation starts with the Treebank, which is performed on three types of source data (DF, SMS/chat and CTS) in three languages (Egyptian Arabic, Chinese and English).

DF source data is manually harvested online by native speakers for each language using a list of topics related to current events and other dynamic events. Collected threads are subsequently triaged down to a selected portion and sentence-segmented for downstream translation and annotation.

SMS and chat source data are collected via live collection platforms and donations. Live-collected messages on free topics are further manually reviewed to exclude any non-target language content, and sensitive or personal information. Chat donations from various sources are screened to discard duplications and split into separate conversations based on a 24 hour-gap. Auto-generated non-text content, such as pictures and emoticons, is removed to facilitate downstream annotation. Resulting data along with metadata including speaker ID, message/conversation ID, date-time are streamlined into LDC’s database.

CTS source data was originally collected for the Arabic and Chinese CallHome and CallFriend program. The data went through Quick Rich Transcript (QRTR) process to produce quick (near-)verbatim, time-aligned transcripts with minimal markup, plus speakerID and SU (sentence-unit) annotations. Collected audio source files are first transcribed by professional transcription agencies, and then go through two passes of quality-control review by LDC junior and senior annotators to assure annotation integrity and freedom from sensitive and personal info.

2.2 Annotation Data

Annotations are performed manually or semi-automatically, providing morphological, syntactic, alignment, and semantic information. Starting with Treebank annotation of source data, other types of annotation are built sequentially from lower-to upper-level layers of linguistic description as shown in Figure 1. The bottom-level Treebank is topped with PropBank and co-reference and then parallel-aligned at the word level. The Treebank and co-reference annotations are performed on Treebank data. Word alignment is built on Treebanks of source and translation data. Source data are only used as input data by Treebank. Tokens extracted from Treebank are fed into propBank, WA, and co-reference annotation pipelines as input data.

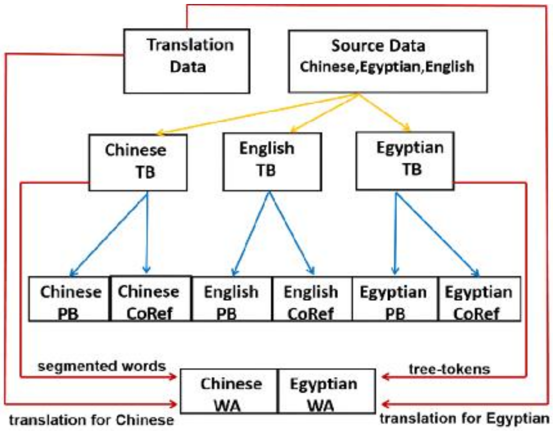


Figure 1: Annotation Pipeline

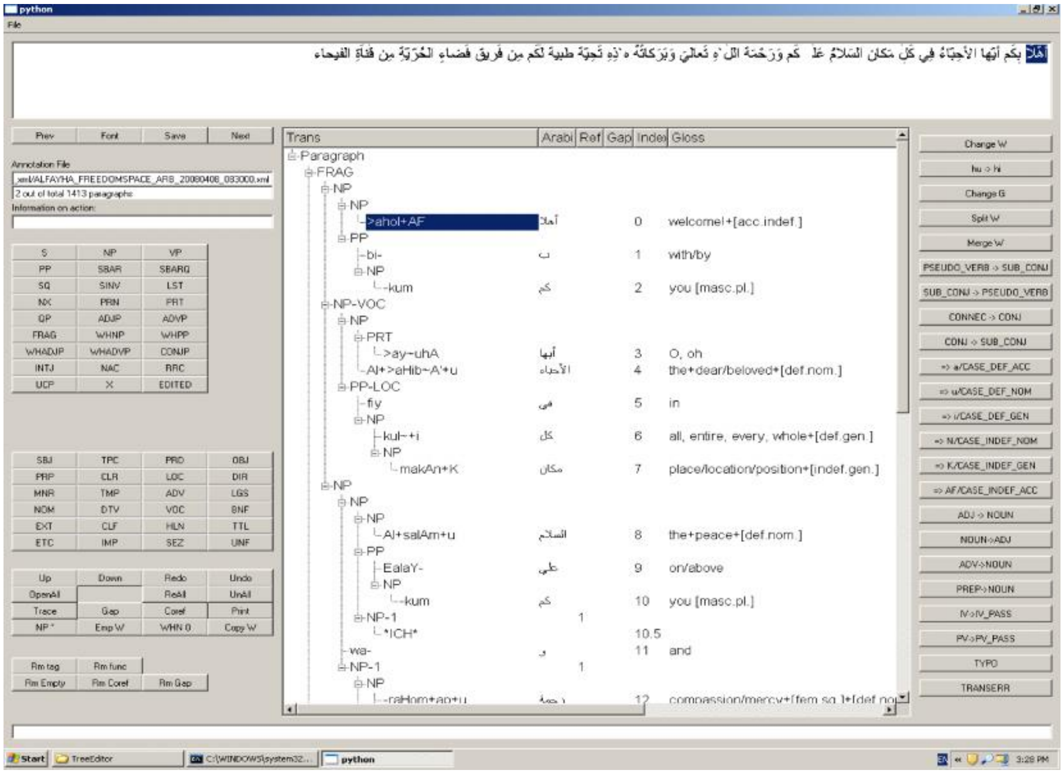
3. Annotation Methodologies

3.1 Treebank

Treebank annotation provides information on word-level tokenization, part-of-speech, and syntactic structures, resulting in a bank of trees representing naturally-occurring linguistic structures. LDC developed a Treebank tool to support morphological and syntactic annotation (Figure 2).

Treebank annotation pipelines include several broad stages:
1. Tokenization of source data, with manual correction
2. Automatic POS tagger or morphological analyzer, with manual correction or selection for consistency with current part-of-speech/morphological annotation guidelines
3. Automatic parser, with manual correction for consistency with Treebank syntactic annotation guidelines
4. Quality control annotation to identify and correct errors

The first stage is the tokenization of source data. For English, the tokenizer hard-codes the handling of abbreviations, numbers, hyphenated words, punctuation, etc. Chinese tokenization is realized through the word segmentation process. Arabic source goes through the SAMA Morphological Analyzer for MSA tokens, and the CALIMA Morphological Analyzer for Egyptian tokens. In the second annotation stage, the tokenized data then goes



أَهْلاً بِكُم أيها الأَحِبَّاءُ فِي كُلِّ مَكانٍ السَّلامُ عَلَ كم وَرَحْمَةُ الل ه تَعالىَ وَبَرَكاتُهُ ه ذِهِ تَحِيَّة طيبة لكَم مِن فَرِيق فَضاءِ الحُرِّيَّة مِن قَناةِ الفيحاء

Annotation File
xml/ALFAYHA_FREEDOMSPACE_ARB_20080408_083000.xml
2 out of total 1413 paragraphs
Information on action:

Trans	Arabi	Ref	Gap	Index	Gloss
Paragraph					
FRAG					
NP					
NP					
>ahol+AF	أهلا			0	welcomel+[acc.indef.]
PP					
-bi-	ب			1	with/by
NP					
--kum	كم			2	you [masc.pl.]
NP-VOC					
NP					
PRT					
>ay~uhA	أيها			3	O, oh
Al+>aHib~A'+u	الأحباء			4	the+dear/beloved+[def.nom.]
PP-LOC					
fiy	في			5	in
NP					
kul~+i	كل			6	all, entire, every, whole+[def.gen.]
NP					
makAn+K	مكان			7	place/location/position+[indef.gen.]
NP					
NP					
NP					
Al+salAm+u	السلام			8	the+peace+[def.nom.]
PP					
EalaY-	على			9	on/above
NP					
--kum	كم			10	you [masc.pl.]
NP-1		1			
ICH				10.5	
wa-	و			11	and
NP-1		1			
NP					
--raHom+ap+u	رحمة			12	compassion/mercy+[fem.sg.]+[def.no

Figure 2: Treebank Annotation Tool

through an automatic POS tagger. The tagger output is then manually corrected to be consistent with the current part-of-speech annotation guidelines and Treebank guidelines. POS tagging divides the text into lexical tokens and includes morphological and morpho-syntactic information. The corrected POS annotated data runs through an automatic parser in the third annotation stage. The parser output is also manually corrected to be consistent with the Treebank syntactic annotation guidelines. Syntactic analysis characterizes the constituent structures of word sequences, providing functional categories for each non-terminal node, and identifies null elements, co-reference, traces, etc.

3.1.1 English Treebank

English POS and Treebank annotation essentially follows the guidelines of Penn Treebank II style, with changes incorporated for the GALE (Global Autonomous Language Exploitation) project and for the BOLT project to account for specific features related to new genres of DF, SMS/Chat and CTS.

English Treebank annotation is performed on two types of source data: English source and English translation data. English source data are directly generated by English user input. English translation data are either translated from the Chinese source or Egyptian Arabic source. Such translation Treebank data is a valuable and important component in constructing parallel aligned Treebank data. English Treebank within BOLT improved annotation quality by adding several rounds of Quality Control (QC) to the annotation process. The first QC process consists of a series of specific searches for approximately 200 types of potential inconsistencies and parser or annotation errors. Any errors found in these searches were hand corrected. An additional QC process then identifies repeated text and structures, and flags non-matching annotations. Identified annotation errors are also manually corrected. A special effort for English translation Treebank is the annotation of alternative translations (Bies et al. 2014). Both literal and fluent translation alternates are annotated for word-level tokenization and part-of-speech, whereas only the fluent translation alternates are annotated as part of the syntactic structure of the tree.

3.1.2 Arabic Treebank

Arabic Treebank started with Penn Arabic Treebank guidelines, enhanced first with the GALE (Global Autonomous Language Exploitation) project and now with the BOLT project (Maamouri et al. 2011). The two efforts of Arabic Treebank within the GALE project are enhancement of Penn style Treebank annotation (Maamouri & Bies, 2010) and the creation of CATiB (Columbia Arabic Tree Bank). With the language type shifting to Egyptian Arabic (Maamouri et al. 2014), additional effort is needed for Egyptian Treebank. First, source Egyptian tokens were automatically converted from Arabizi (or Romanized/Latin characters) script to Arabic script (Bies et al. 2014). The automated results were manually corrected before Treebanking. Second, for POS annotation, a new “wildcard” feature was introduced to handle tokens with solutions in neither SAMA nor CALIMA to allow annotators to supply annotation for a stem that is not in either analyzer. Third, to enhance syntactic annotation quality, annotators went through a stage of annotation with the help of diagnostic QC searches to catch potential patterns of annotation errors.

3.1.3 Chinese Treebank (CTB)

Treebanking approaches for Chinese source originated with Penn Chinese Treebank project (Xue et al. 2005) and OntoNotes (Weischedel et al. 2013). With BOLT Chinese Treebank, a new annotation workflow was implemented for an annotation efficiency increase of 30%. The previous two distinctive annotation stages (POS tagging and Treebanking) were decomposed into five self-contained steps: sentence boundary detection, word segmentation/POS tagging, constituent grouping, functional category and empty category annotation, and post-processing and validation. This new workflow helps to classify annotators, reduce cognitive load of annotators, and enlarge qualified annotator pool for quick annotation. This new workflow also allows cross-checking between different layers of annotation, thus enhancing annotation quality. Chinese Treebank introduced new tags for handling challenging informal genre features, such as typographical errors, non-speech elements, incomplete utterances, fillers, embedded utterances, etc.

3.1.4 Treebank Data Structure

Treebank annotation results are stored in .tree files in the basic bracketed format of a pair of unlabeled parentheses, with one segmented sentence per line, as shown in the following sample output. For sentence segments containing more than one sentence, the pair of unlabeled parentheses will contain all necessary top level sentences. It is possible for one line in .tree files to include more than one tree.

((S (NP-SBJ (PRP I)) (VP (MD 'll) (VP (VB post) (NP (NP (NNS highlights)) (PP (IN from) (NP (DT the) (NN opinion) (CC and) (NNS dissents)))) (SBAR-TMP (WHADVP-9 (WRB when)) (S (NP-SBJ (PRP I)) (VP (VBP 'm) (ADJP-PRD (JJ finished)) (ADVP-TMP-9 (-NONE- *T*)))))) (. .)))

POS annotation is stored in .pos files, where POS tags are attached to tokenized/segmented word units, each sentence per line, as shown in the following example.

好_VA 的_SP 呗_SP ，_PU 来_VV 的话_SP 打_VV 我_PN 电话_NN 就_AD 可以_VV ，_PU 或者_CC 报_VV 我_PN 名字_NN ，_PU 我_PN 定_VV 了_AS 包厢_NN

3.2 Word Alignment

3.2.1 Alignment Approach

Word alignment guidelines are developed based on guidelines for the Blinker and ARCADE projects, enriched during GALE by adding tagging guidelines (Li et al. 2010), and further enhanced to tackle new genres and language features for the BOLT project (Li et al. 2012). The alignment annotation involves a process of 2-pass annotation plus one round of cross-file checking. The initial alignment by junior annotators goes through quality control by senior annotators, and is then followed by a cross-file check by lead annotators for consistency. The word alignment tool is developed by LDC (Figure 3), allowing annotators to align source and translation words as well as to label both alignment links and individual words.

Alignment is performed on two pairs of languages: Chinese-English and Egyptian-English. Chinese alignment is performed at two levels: character-level and CTB-token level. Whitespace is inserted as the character delimiter for character-level alignment. Character-level alignment is manually performed. CTB-token alignments are automatically generated from character-level alignments. The Chinese word alignment focuses on aligning and tagging 8 different types of links, 14 types of unmatched words, and all instances of Chinese 的 (DE) when it is used as a function word. The Chinese tagging task is designed to capture both functional and syntactic information of the unaligned words and aligned links, such as “has completed” is aligned to “完成了”, where the alignment link is tagged as grammatically-inferred link as this alignment has function words "has" and "了", and these two function words are further tagged as "tense marker". Egyptian Arabic word alignment has fewer link types and word tags. Four types of links are designed for alignment, and only one tag is used for tagging unaligned function words which are attached to content words.

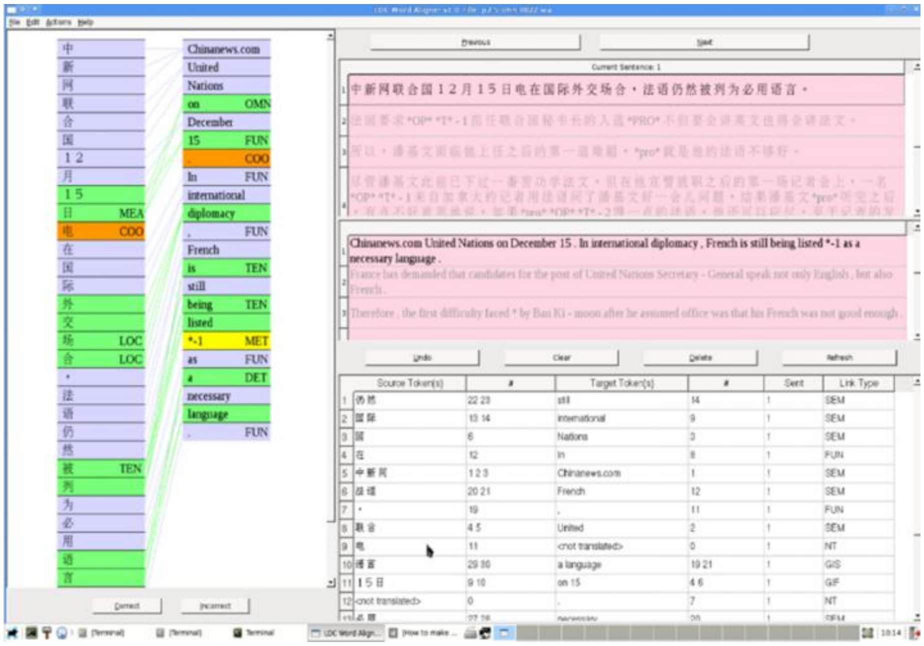
3.2.2 Alignment Data Structure

The alignment annotation is stored in .wa file. The format of the alignment file is similar to the GIZA++ word alignment format, but with some enhancements. Each line contains a list of space delimited alignments for the corresponding sentence. The alignments appear in no particular order. Each alignment is in the format of s-t(linktype), where s and t are a list of comma delimited source and translation token IDs respectively. S or t can be empty indicating an unaligned token. Additionally, each token number may be optionally followed by a tag in square brackets. Chinese/Egyptian token numbers always appear before the hyphen and English tokens always after no matter which language is the source or translation. In the following alignment example, source token 22 is linked to translation tokens 25, 26, 27, and 28. Translation tokens 25 and 28 are tagged as OMN and POS respectively. The link type is GIS.

22-25[OMN],26,27,28[POS](GIS)

3.2.3 Alignment Data Use

One peculiar feature about WA data structure is that users can flexibly tailor data for different MT models. Some MT models prefer heavy linguistic rules while others prefer less linguistic input. For instance, if a MT model prefers less vocabulary size and would want to have all the unmatched



1 中新网联合国12月15日电在国际外交场合，法语仍然被列为必用语言。

2 法国要求*OP* *T*-1担任联合国秘书长的人选*PRO*不但要会讲英文也得会讲法文。

3 所以，潘基文面临他上任之后的第一道难题，*pro*就是他的法语不够好。

4 尽管潘基文此前已下过一番苦功学法文，但在他宣誓就职之后的第一场记者会上，一名*OP* *T*-1来自加拿大的记者用法语问了潘基文好一会儿问题，结果潘基文*pro*听完之后，有点不好意思地说，如果*pro* *OP* *T*-2慢一点的话，他还可以应付，至于记者的发

1 Chinanews.com United Nations on December 15 . In international diplomacy , French is still being listed *-1 as a necessary language .

2 France has demanded that candidates for the post of United Nations Secretary - General speak not only English , but also French .

3 Therefore , the first difficulty faced * by Ban Ki - moon after he assumed office was that his French was not good enough .

	Source Token(s)	#	Target Token(s)	#	Sent	Link Type
1	仍然	22 23	still	14	1	SEM
2	国际	13 14	international	9	1	SEM
3	国	6	Nations	3	1	SEM
4	在	12	In	8	1	FUN
5	中新网	1 2 3	Chinanews.com	1	1	SEM
6	法语	20 21	French	12	1	SEM
7	。	19	,	11	1	FUN
8	联合	4 5	United	2	1	SEM
9	电	11	<not translated>	0	1	NT
10	语言	29 30	a language	19 21	1	GIS
11	1 5 日	9 10	on 15	4 6	1	GIF
12	<not translated>	0	.	7	1	NT
13	必用	27 28	necessary	20	1	SEM

Figure 3: Word Alignment Tool

lexical units to be left unaligned, the data can be automatically pre-processed to detach all tagged unaligned words by making use of word tags. This is possible because all unaligned words in annotation data are tagged. It's also possible for users to manipulate word tags differently for producing different annotation for their model. WA tagging and alignment data structure provide users with more affordable alternative research approaches because a quick and automatic pre-processing of data is far less expensive than re-annotating data with a different annotation scheme.

3.3 PropBank

3.3.1 PropBank Approach

PropBank provides a layer of semantic annotation on top of the phrase structure of Treebank. PropBank annotation is supported with the Jubilee interface implemented by the University of Colorado (Figure 4), where any node in the tree can be selected and assigned tags.

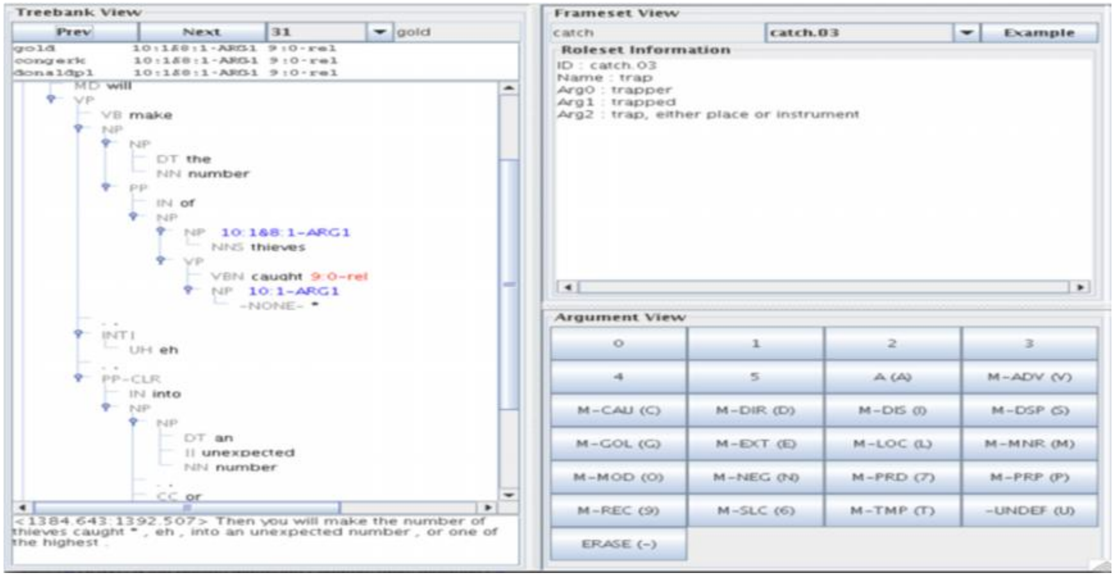
Each predicate verb in a tree is annotated with a sense id and the semantic role labels for its arguments. This annotation aims to provide consistent semantic role labels across different syntactic realizations of the same verb, to assign functional tags to all non-core arguments of the verb, such as manner, location, and temporal, and to find antecedents for 'empty' arguments of the predicates, such as [*] in the sentence, "I made a decision [*] to leave", creating co-reference chains for empty categories, relative clauses, and reduced relatives.

Predicate argument structure annotation is carried out in two phases. In the first phase, a frame file for a predicate is created by examining all instances of the predicate in the Treebank data and distinguishing two or more senses, which are called Framesets or Rolesets. In the second phase, the predicate argument structure of all instances of the predicate are annotated, using the Frame File as a reference. The arguments of each predicate receive an argument label in the form of ArgN, where N is an integer between 0 and 6. These numbered arguments represent core arguments that are defined in relation to the predicate. Each core argument plays a unique role with regard to the predicate.

Core arguments are as consistent as possible with respect to thematic roles. Arg0 is used for the most agentive role a given predicate can take. Arg1 is used for the proto-patient, or most patient-like argument. Arg2 is most often used to mark a beneficiary, Arg3 is most often used to show a start point, and Arg4 is most often used for the end point. Args2-4 are less consistent, as not all verbs with more than 2 core roles require a start/end point role or a beneficiary, so these are used in other ways as dictated by a given predicate.

Due to differing grammatical structures, each language faces the challenge of defining predication. Verbs alone are not sufficient to project sentimental semantics in all environments. In Chinese, English, and Arabic, specific constructions involving nouns and adjectives can project semantics, overriding the verb. In other instances, an eventive noun and a light verb work can together to form a light verb construction to convey meaning.

Chinese PropBank focuses on annotating the predicate-argument structure of all verbs and their nominalizations, using the guidelines and frame files developed for the Chinese Propbank (CPB) (Xue and Palmer 2009). Chinese PropBank tackles some special grammatical constructions where arguments of certain verbs are systematically dislocated from their canonical positions, such as BA(把)-



```
gold        10:1&0:1-ARG1 9:0-rel
congerk     10:1&0:1-ARG1 9:0-rel
donaldp1    10:1&0:1-ARG1 9:0-rel
```

```
      MD will
  VP
      VB make
      NP
          NP
              DT the
              NN number
          PP
              IN of
              NP
                  NP  10:1&8:1-ARG1
                      NNS thieves
                  VP
                      VBN caught 9:0-rel
                      NP  10:1-ARG1
                          -NONE- *
      , ,
      INTJ
          UH eh
      , ,
      PP-CLR
          IN into
          NP
              NP
                  DT an
                  JJ unexpected
                  NN number
              , ,
              CC or
```

<1384.643:1392.507> Then you will make the number of thieves caught * , eh , into an unexpected number , or one of the highest .

catch

catch.03

ID : catch.03
Name : trap
Arg0 : trapper
Arg1 : trapped
Arg2 : trap, either place or instrument

Figure 4: PropBank Annotation Tool

construction, BEI(被)-construction, and You(由)-construction. The challenge of annotating nominalized predicates is that certain nouns in Chinese are not true nominalizations even though they have a verbal counterpart that shares the same morphological form. In addition, not all modifiers of a nominalized predicate can be considered as its arguments or adjuncts. Some modifiers only occur with the nominal form of a predicate but never with its corresponding verb form. In this case they are not considered to be arguments of the predicate

Arabic PropBank does not to distinguish between nouns and adjectives, since both are predicative. Arabic PropBank annotates adjectives, nouns, and light verb constructions. A major challenge for Arabic was with special features of the Egyptian dialect. It was sometimes very difficult to decide if an MSA word form in the dialect had an equivalent meaning or a slightly different meaning. Additionally, as a pro-drop language, Arabic poses special issues for co-reference. Co-reference chains creation problems were initially solved by manual annotation but are now able to be created during post-processing.

The English PropBank has focused on expanding predicate annotation beyond the verb and is now annotating verbs, eventive nouns, adjectives, and light verb constructions (Bonial et al. 2014). In English, light verbs are semantically bleached verbs. The argument structure for light verb and non-light verb instances are different. In light verb constructions, the real predicate is generally the nominalized predicate that the light verb supports. A major focus for English PropBank has been to unify FrameFiles across these different parts of speech. This means that the frame used for 'bathe' is always identical to that used for 'bath'. The goal of this expansion is to provide event semantic representations for the entire sentence, specifically pieces most often missed when looking solely at verbs, consistent with AMRs (Banarescu et al 2013)

3.3.2 PropBank Data Structure

PropBank annotation results are stored in two types of files: Frame Files (.xml) and PropBank files (.prop). The frame file is in XML format. At the top level, each verb frame file is associated with a set of "framesets". A frameset contains a set of roles and a set of frames. Each frameset has an identifier, and associates the verb with a verbclass. A frame consists of a set of tuples of mapping and example elements. A mapping is an ordered list of mapitems, with a verb included. The English frame file in XML format includes content of the unification of noun, adjective and verb predicates.

The .prop file contains PropBank instances (one instance per line). Each line represents the argument structure of one predicate instance and consists of space-separated columns, as shown in the following sample output:

bolt6/SMS_CMN_20130706.0001.su.fid 15 22 gold 定-v 定.01 ----- 21:1-ARG0 22:0-rel 24:1-ARG1

3.4 Co-reference

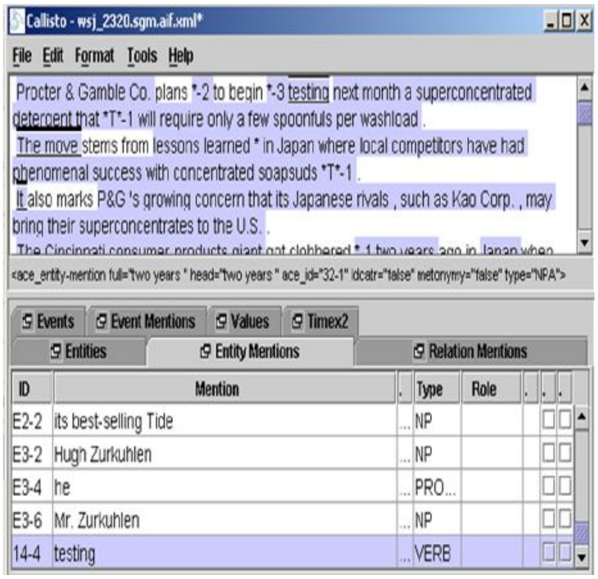
3.4.1 Co-reference Approach

The purpose of the co-reference annotation task is to fill in all of the connections between specific mentions in the text that refer to the same entities and events in the discourse context. Co-reference here is limited to noun phrases (including proper nouns, nominals, pronouns. and null arguments), possessives, proper noun pre-modifiers, and verbs.

The annotation distinguishes between two types of co-reference: Identity (IDENT) and Appositive (APPOS). IDENT chains are used to mark cases of anaphoric co-reference. An anaphoric co-reference refers to links which are not directly signalled in the syntax between pronominal, nominal and named mentions of specific referents. In order to mark IDENT co-reference, there must be a specific mention, usually pronominal, named, or definite nominal. It does not include entities that are only mentioned as generic, underspecified or abstract. Pre-modifiers that are proper nouns are eligible for co-reference.

Appositives are signalled in the syntax by an appositive construction, which contains a noun phrase that is modified by one or more immediately-adjacent noun phrase(s), separated by only a comma, colon, or parenthesis. The two parts of an appositive, the head (noun phrase that points to an object or concept in the world) and one or more attributes of that referent, are both annotated. The single span containing the entire appositive construction can also in turn be linked as part of an IDENT chain.

Co-reference annotation is performed with the Callisto tool developed by MITRE (Figure 5), where verbs, pronouns and proper pre-modifiers are added when they are co-referent with an NP, and they are labelled with entity mentions types.



Procter & Gamble Co. plans *-2 to begin *-3 testing next month a superconcentrated detergent that *T*-1 will require only a few spoonfuls per washload .
The move stems from lessons learned * in Japan where local competitors have had phenomenal success with concentrated soapsuds *T*-1 .
It also marks P&G 's growing concern that its Japanese rivals , such as Kao Corp. , may bring their superconcentrates to the U.S. .
The Cincinnati consumer products giant got clobbered * 1 two years ago in Japan when

<ace_entity-mention full="two years " head="two years " ace_id="32-1" idcatr="false" metonymy="false" type="NPA">

ID	Mention	.	Type	Role	.	.	.
E2-2	its best-selling Tide	...	NP			□	□
E3-2	Hugh Zurkuhlen	...	NP			□	□
E3-4	he	...	PRO...			□	□
E3-6	Mr. Zurkuhlen	...	NP			□	□
14-4	testing	...	VERB			□	□

Figure 5: Co-reference Annotation Tool

3.4.2 Co-reference Data Structure

Co-reference annotation output data is stored in XML format. Each line of the xml file lists the tokens from one sentence in the underlying Treebank release, and XML <COREF> brackets are wrapped around each mention of a coreferent entity. In the following examples, the ID attribute number tells which entity this mention belongs to. "IDENT"-type tags are used for normal linguistic co-reference. "APPOS" tags are used to mark the elements in an appositive construction.

And <COREF ID="1" TYPE="IDENT">Iran</COREF> will never dare *PRO* nuke <COREF ID="16" TYPE="IDENT">US</COREF>, not even *PRO* using terrorists.

4. Data Consistency and Integrity

To assure annotation consistency, various quality control strategies (see Section 3) are in place for boosting annotator training and annotation quality. With annotator training, for instance, new Treebankers are trained on documents already annotated in the CTB until their agreement with existing annotation reaches 90% or above before they start productive annotation. Word alignment annotators go through guidelines training, demo-annotation training, and accuracy testing before production. PropBank and co-reference use annotators with experience from previous annotation projects, particularly the GALE OntoNotes project, and any new annotators have been trained on a set of trial data until they reach an adequate level of consistency before they start production-level annotation. To assure annotation quality through the entire production process, annotators frequently meet to discuss questions and pop-up issues encountered during annotation. These meetings also serve as trainings for new annotators and as a refresher course for seasoned annotators.

In addition to rigorous training process, annotation quality is reinforced by multiple rounds of annotation. All types of annotation undergo the first pass of annotation by junior annotators, followed by the second pass annotation by senior annotators and the final quality-control pass by lead annotators. Automatic and semi-automatic checks are introduced to facilitate annotators in identifying and fixing annotation errors within and across annotation files. Annotation consistency rate varies in terms of different types of annotation and different languages/dialects. For instance, propBank annotation is particularly challenging for Egyptian Arabic (82% inter-annotator rate) due to its dialect features, while the Chinese propBank IAA is relatively higher (91.2%).

To assure data consistency across annotation types, annotation data is designed to be structured and stored using identical tokens, identical sentences and files, and identical filenames for all types of annotation data. Validations and sanity checks are performed to ensure this data identicalness and integrity at each annotation level and across annotation datasets. Cross-corpus data consistency and integrity allow users to conveniently plug in any layer or all layers of annotation for linguistic analysis and system training.

5. Conclusion

Unstructured and informal internet data result in language irregularities and ambiguities, making it difficult for traditional systems to process and digest. Non-standard lexical items and syntactic patterns lead to unintentional errors, dialectal variation, conversational ellipsis, topic diversity, and creative use of language and orthography. Annotations produced at various language levels in various genres can help to explain this widespread variation.

Tagging unstructured information at various linguistic levels eliminate ambiguities for enhancing translation automation and information retrieval. The large multi-layer, multi-lingual, and multi-genre corpus produced at LDC bridged the gap between unstructured and structured data, uncovering hidden linguistic insights and providing a rich resource for training and evaluating NLP systems. Previously distributed to BOLT performers, the annotation data is now being prepared for broader distribution to LDC members and non-member licensees via publication in the LDC catalogue.

6. Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2013) Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria.

Bies, A, Mott, J., Kulick, S., Garland, J., and Warner, C. (2014). Incorporating Alternate Translations into English Translation Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland

Bies, A, Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. *Arabic Natural Language Processing Workshop, EMNLP*. Doha, Qatar.

Bonial, C., Bonn, J., Conger, K., Hwang, J., and Palmer, M. (2014). PropBank: Semantics of New Predicate Types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland

Castor, A., Pollux, L.E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1), pp. 37--53.

Cotterell, R. and Callison, B.C., (2014). A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland.

Jha, M., Andreas, J., Thadani, K., Rosenthal, S and McKeown, K., (2010). Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment. In *Proceedings of NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*

Li, X., Ge, N., Grimes, S., Strassel, S. M. and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta.

Li X., Strassel, S., Grimes, S., Ismael, S., Maamouri M., Bies, A. and Xue, N. (2012). Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures. In *Proceedings of 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N. and Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland

Maamouri M. and Bies A. (2010). The Penn Arabic Tree Bank. In Ali Farghaly (Ed.), *Arabic Computational Linguistics*. CSLI Studies in Computational Linguistics.

Maamouri M. Bies A, Kulick, S., Habash, N., Reem Faraj and Ryan Roth. (2011). Arabic Treebanking. In Joseph Olive, Caitlin Christianson and John McCary (Eds.), *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Owoputi, O., O'Connor, B., Dyer, C., Gimpely, K., Schneider, N. and Smith, A.N. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 380-391

Read, J., Flickinger, D., Dridan, R., Oepen, S., and Øvrelid, L. (2012). The WeSearch Corpus, Treebank, and Treecache -- A Comprehensive Sample of User-Generated Content. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Istanbul, Turkey

Seth Kulick, Ann Bies, Justin Mott, Mohamed Maamouri, Beatrice Santorini, Anthony Kroch. 2013. "Using Derivation Trees for Informative Treebank Inter-Annotator Agreement." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2013)*. 2013.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, Mi., Bachouti, M.E., Belvin, R., Houston, A. (2013). OntoNotes Release 5.0. https://catalog.ldc.upenn.edu/LDC2013T19

Xue, N and Palmer, M. (2009). Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143-172.

Xue, N., Xia, F. Chiou, Fu., and Palmer M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207-238.