Corpus Development for the ACE (Automatic Content Extraction) Program

Alexis Mitchell & Stephanie Strassel
Linguistic Data Consortium
{amitche0,strassel}@ldc.upenn.edu

Introduction

Automatic Content Extraction

- Designed to support automatic processing of source language text data, including classification, filtering, and selection based on meaning of the source data
- Ultimate goal
  - Development of technologies that automatically detect and characterize this meaning
- ACE applications will
  - Maintain a database of what is happening in the world
  - Ideally, this will be in terms of who is doing what, where and when
  - Database will maintain pointers into source data
- Research objectives – detection and characterization of
  - Entities (Phase 1)
  - Relations (Phase 2)
  - Events (Phase …)

ACE Timeline

- Pilot Phase: Entity Detection & Tracking (EDT)
  - Test initial EDT guidelines through small multi-site pilot annotation
    - 15K words triply annotated
  - May, November 2000 Evaluations
- Phase 1 adds metonymy, generics (EDT+)
  - Further refinement of EDT guidelines
  - Multi-site annotation
    - Training/development data triply annotated
    - Evaluation data annotated by LDC
  - February 2002 evaluation

ACE Timeline

- Phase 2: Relation Detection & Characterization (RDC)
  - LDC as sole annotation site
  - LDC role in annotation spec development
  - September 2002 evaluation
    - Shared evaluation with Evidence Extraction (EE) community
- Future: Modifications, enhancements to EDT, RDC tasks
  - Continued synergy with EE community
  - Possible reworking of sticky EDT issues
  - Phase 3: Event detection & characterization

Data Resources

- English newswire, BN transcripts, newspaper
  - Sites work with OCR newspaper output
- Phase 1: Entity Detection & Tracking
  - 15K words pilot data
  - 180K training/development data
  - 45K evaluation data
- Phase 2: Relation Detection & Tracking
  - Entire ACE Phase 1 corpus *plus*
  - 45K new evaluation data
  - 50K new data from Evidence Extraction & Link Detection (EELD) community (domain-specific)
- Annotated corpora slated for release as regular LDC publications

Annotation Task

EDT Overview

- Annotators identify all entities of type
  - Person — *Bush, he, the President*
  - Organization — *Linguistic Data Consortium*
  - Facility — *Alfredo Kraus Auditorium*
  - Location — *the Hudson River*
    - Geo-Political Entities (GPE) with Role
      - GPE.PER if referent is population of GPE — ***Cubans** protested…*
      - GPE.LOC if referent is territory of GPE — *the **U.S.** heartland*
      - GPE.ORG if referent is government of GPE — ***Iraq** agreed…*
      - GPE.GPE if referent is whole GPE — ***U.S.** leader*

Mentions

Mention Extent

- Maximal extent of NPs

  {Mrs. Adamson, whose cheerful, under-five-foot presence is strengthened by soft blue eyes and spun-silver hair}

- Head of NP

  {**[Mrs. Adamson]**, whose cheerful, under-five-foot presence is strengthened by soft blue eyes and spun-silver hair}

- Nested Mentions

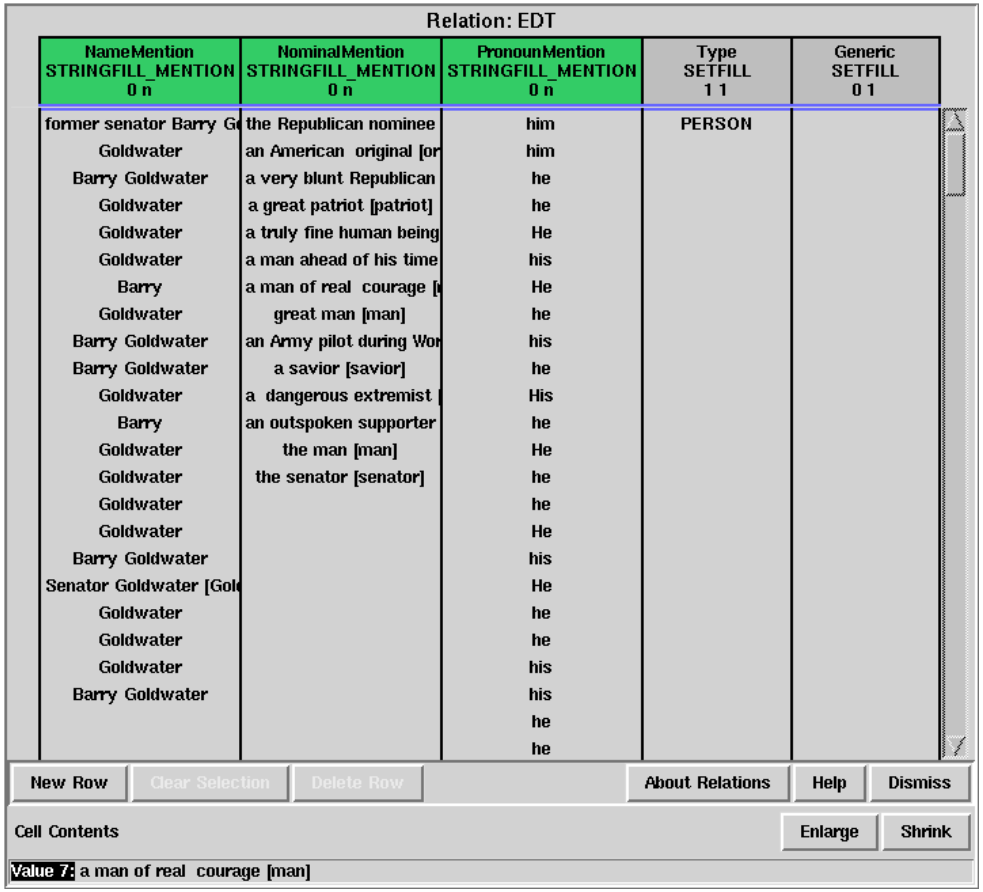
  {**[Washington]** lawyer **[Vernon E. Jordan Jr.]**, {**[one]** of {the **[president]**'s closest **[advisers]**}}}

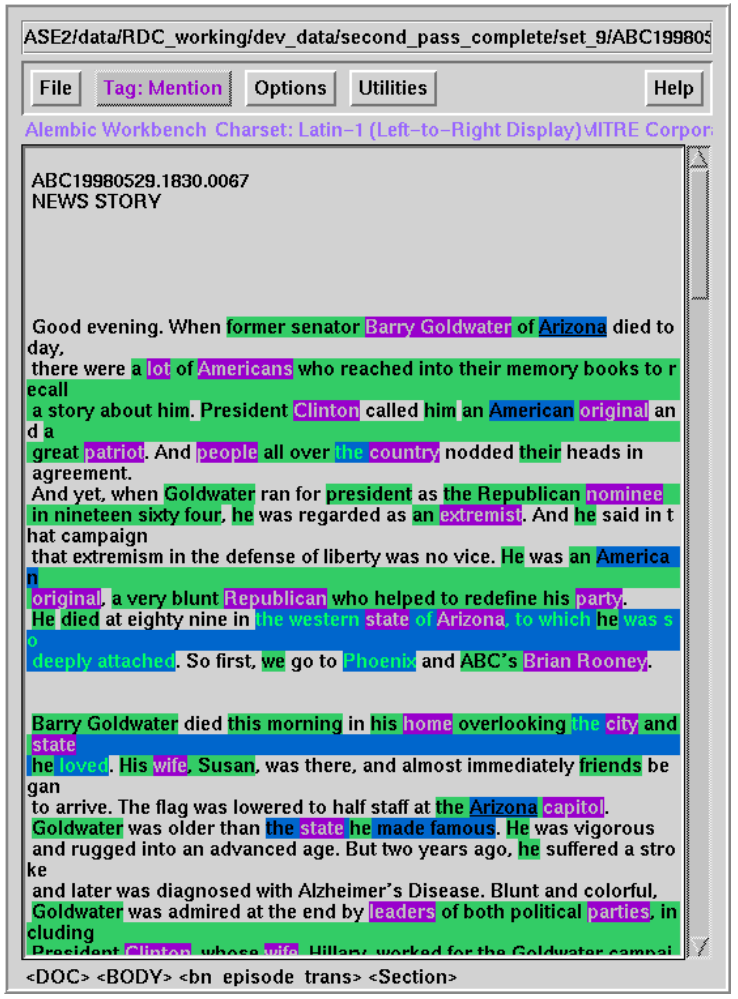
Mention Type

- Proper — **France**, The **[Washington Post]**, **Kenneth Starr**
- Common — the house **[painters]**, the **[hospital]**, a suburban **[community]**
- Pronominal — **her**, **our**, **you**, **its**, **one**

Coreferencing Mentions

- Coreference all mentions of same entity within document





Entity Class (generic/specific)

- Generic/Specific
  - Specific Instance: *Reporters* covering the trial…
  - Generic Class: *Reporters* don't reveal sources.
- Co-Indexing Generics

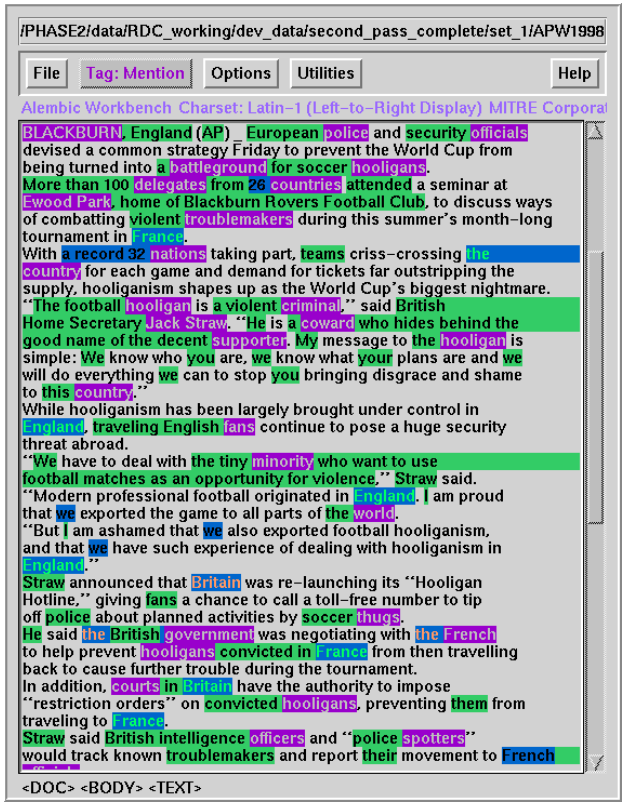
Relation: EDT

	NameMention STRINGFILL_MENTION 0 n	NominalMention STRINGFILL_MENTION 0 n	PronounMention STRINGFILL_MENTION 0 n	Type SETFILL 1 1	Generic SETFILL 0 1
22		The football hooligan [ho a violent criminal [crimin a coward who hides beh the hooligan [hooligan] the tiny minority who w soccer thugs [thugs] troublemakers those who cause trouble the hooligan [hooligan] soccer hooligans [hooliga violent troublemakers [t	He you your you their	PERSON	TRUE

New Row | Clear Selection | Delete Row | About Relations | Help | Dismiss

Cell Contents | Enlarge | Shrink

Value 6: soccer thugs [thugs]



Metonymy

- Identify and characterize metonymy
  - GPEs
    - **Beijing** will not continue sales of anti-ship missiles to Iran.

      {[GPE.GPE: literal] [GPE.ORG: intended] **Beijing**}
  - Organizations and Facilities
    - A few hundred ethnic Albanians laid a black wreath at the gate of the Yugoslavian **embassy**.

      {[ORG:literal] [FAC:intended] the Yugoslavian **embassy**}.
  - Sports Teams
    - **Brazil** made it to the final round of the World Cup.

      {[GPE.GPE: literal] [ORG:intended] **Brazil**}

RDC Overview

- Builds on EDT Annotation
- Annotators establish relations between pairs of entities
  - Explicit relations
    - *President Clinton* was in *Washington* today.
  - Implicit relations
    - In what appeared to be an effort to divert some flack away from *Zhu*, Hu Jintao, another member of *the Standing Committee*, is leading the working committee nominally in charge of devising the streamlining plan.
  - Limited set of relation types and subtypes
    - Combination of ACE- and EE-inspired relations

Relation Types

- 5 Relation Types
  - AT
    - President Bush gave a speech in New Jersey last month.
  - PART
    - Dallas, TX
  - ROLE
    - US government spokesperson
  - NEAR
    - The train station is right outside Media.
  - SOCIAL (SOC)
    - Joe called his cousin the other day.

Relation Subtypes

- AT
  - Located — George Bush gave a speech in New Jersey.
  - Based-In — The US company has many branches worldwide.
  - Residence — Hillary Clinton moved to New York last year.
- PART
  - Part-Of — Philadelphia, Pennsylvania
  - Subsidiary — Microsoft's accounting office
- NEAR
  - Relative-Location — The park is two blocks from Walnut Street.

Relation Subtypes (Cont.)

- ROLE
  - Management — the **CEO** of Microsoft
  - General-Staff — Mr. Smith, a **programmer** at Microsoft
  - Member — the permanent UN member **countries**
  - Citizen-Of — Jean-Luis is French.
  - Owner — Joe has decided to remodel his house.
  - Founder — the **founder** of the **University of Pennsylvania**
  - Affiliate — Philadelphia is the sister **city** of Florence, Italy.
  - Client — Bill Clinton's **lawyer**

Relation Subtypes (Cont.)

- SOC
  - Parent — Joe's **father** retired last week.
  - Spouse — **Joe and Sarah got married last night.**
  - Sibling — Joe's **brother** ran a marathon.
  - Grandparent — Joe's **grandmother** is 100 years old.
  - Other-Relative — Joe and his cousin went fishing.
  - Other-Personal — Bill is Joe's **neighbor**.
  - Associate — Mary and her teammates
  - Other-Professional — Schwartz's **students**

Relations and EDT Types

ARG1 \ ARG2	FAC	GPE	LOC	ORG	PER
FAC	PART.Part-of AT.Located	AT.Located NEAR.Relative-location PART.Part-of	AT.Located NEAR.Relative-location	PART.Part-of	
GPE	NEAR.Relative-Location ROLE.Owner	NEAR.Relative-location PART.Part-of ROLE.Affiliate ROLE.Member ROLE.Client ROLE.Management ROLE.Other	AT.Located NEAR.Relative-location PART.Part-of	ROLE.Member ROLE.Client ROLE.Owner ROLE.Affiliate ROLE.Other	ROLE.Member ROLE.Client ROLE.Affiliate ROLE.Other
LOC	NEAR.Relative-location PART.Part-Of	NEAR.Relative-location PART.Part-of	NEAR.Relative-location PART.Part-of		
ORG	ROLE.Owner AT.Located	AT.Located AT.Based-In PART.Subsidiary PART.Part-of ROLE.Affiliate PART.Other ROLE.Member ROLE.Client ROLE.Other ROLE.Management	AT.Located AT.Based-In ROLE.Owner	PART.Subsidiary PART.Other PART.Part-of ROLE.Member ROLE.Affiliate ROLE.Client ROLE.Other ROLE.Management ROLE.Founder	ROLE.Member ROLE.Client ROLE.Other
PER	AT.Located AT.Residence NEAR.Relative-location ROLE.Owner ROLE.Founder ROLE.General-staff ROLE.Management ROLE.Other	AT.Located AT.Residence NEAR.Relative-location ROLE.General-staff ROLE.Member ROLE.Management ROLE.Citizen-of ROLE.Client	AT.Located AT.Residence NEAR.Relative-location ROLE..Owner	AT.Located ROLE.General-staff ROLE.Member ROLE.Management ROLE.Owner ROLE.Founder ROLE.Client ROLE.Affiliate ROLE.Other	SOC.Parent SOC.Sibling SOC.Spouse SOC.Grandparent SOC.Other-relative SOC.Other-personal SOC.Associate SOC.Other-professional ROLE.Member ROLE.Client ROLE.Affiliate

Relation Coreference

Coreference

- Equivalent values in Class, Type, Subtype, Entity of ARG1, and Entity of ARG2 fields
- RDCID-2
  - AT.Residence relation

"He died at eighty nine in the western state of Arizona…

"former senator Barry Goldwater of Arizona"

Relation: RDC

	RelationID STRING_RDCID 1 1	RelationClass SETFILL 1 1	Type SETFILL 1 1	Subtype STRING_RDCSUBTYPE 1 1	Argument1 STRINGFILL_RDCARG 1 1	Argument2 STRINGFILL_RDCARG 1 1	Time STRINGFILL_REL_TIME 0 n
1	RDCID-1	EXPLICIT	ROLE	Management	former senator Barry G	Arizona [EDT-2, GPE]	[RELINSTSLOT tag]
5	RDCID-2	EXPLICIT	AT	Residence	He [EDT-1, PER]	the western state of A	[tag]
2	RDCID-2	EXPLICIT	AT	Residence	former senator Barry G	Arizona [EDT-2, GPE]	
4	RDCID-3	EXPLICIT	ROLE	Member	his [EDT-1, PER]	his party [EDT-8, ORG	
3	RDCID-3	EXPLICIT	ROLE	Member	the Republican nominee	Republican [EDT-8, OR	[tag]
15	RDCID-4	EXPLICIT	ROLE	General-Staff	Brian Rooney [EDT-11	ABC News [EDT-12, O	
6	RDCID-4	EXPLICIT	ROLE	General-Staff	ABC's Brian Rooney [E	ABC [EDT-12, ORG]	
7	RDCID-5	EXPLICIT	AT	Located	Barry Goldwater [EDT-	his home overlooking th	[tag]
8	RDCID-6	EXPLICIT	ROLE	Owner	his [EDT-1, PER]	his home overlooking th	
9	RDCID-7	EXPLICIT	AT	Located	his home overlooking th	the city [EDT-10, GPE	
10	RDCID-8	EXPLICIT	SOC	Spouse	His [EDT-1, PER]	His wife, Susan [EDT-	

New Row | Clear Selection | Delete Row | About Relations | Help | Dismiss

Cell Contents | Enlarge | Shrink

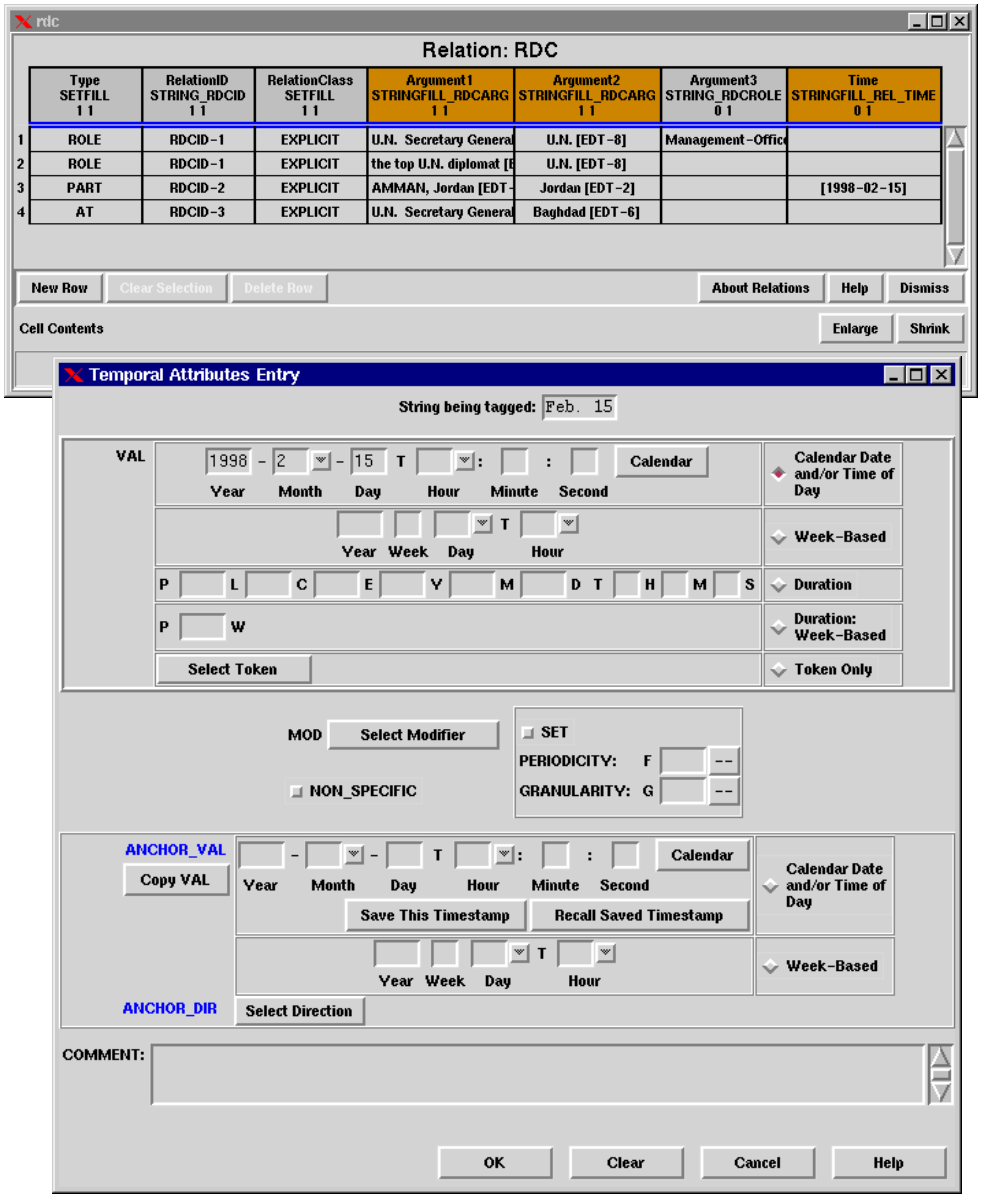
Value 1: ROLE

Temporal Relation Attributes

- Builds on TIMEX2 tagging
- Annotate temporal attributes of explicit relations *only*
  - Specific, Absolute
    - Specific calendar values
      - Blair's visit to China in 1998.
    - Calendar values in relation to anchor value (date of news story)
      - The inspectors left the site last week.
  - General, Relative
    - Indicated by tense of finite verb that heads predicate of relation
    - Relation holds *before*, *as-of*, or *after* anchor value
      - Bush visited Russia.
  - General, Unspecified
    - Point in time or duration of time
    - Without absolute or relative temporal value
      - The inspector's appearance in Baghdad at the appropriate time…
      - The fugitives remained in the compound for eight days.
- Time attributes most frequent with
  - AT relation types
  - PERSON (Arg1) entity types

Tagging Temporal Attributes

- For relations with temporal component,
  - Look for pre-existing TIMEX2 tag
    - Tag it as **attribute** of relation mention pair
  - If no prior TIMEX2 tag, create new REL_TIME tags
- One relation can have multiple timestamps
  - She was in Las Vegas in May and again in June.
- Time range expressions require start and end point annotation in TIME_RANGE tag



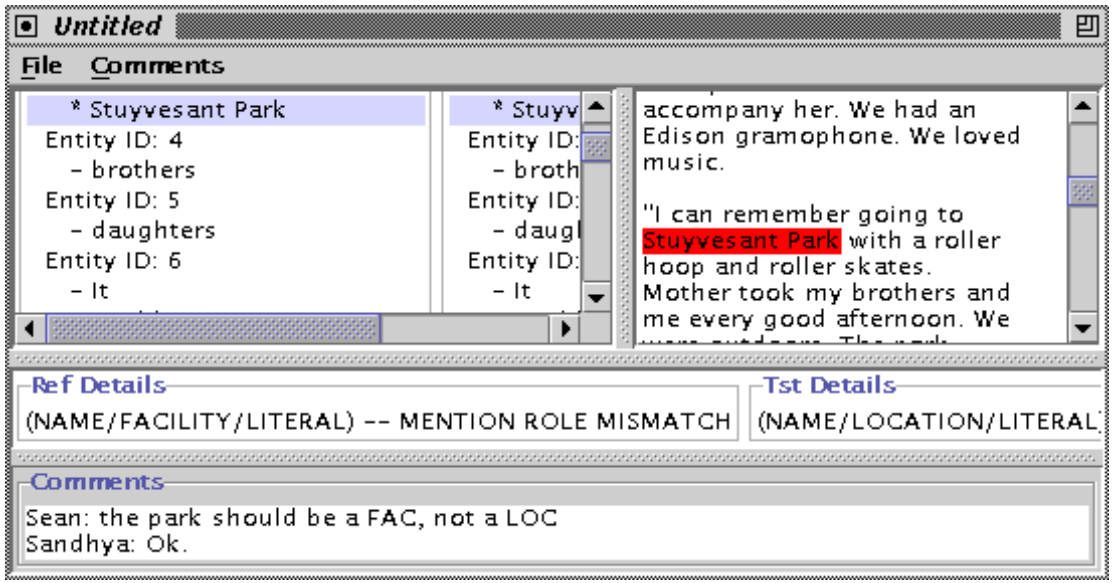
Temporal annotation issues

- Open-ended time attributes
  - *Kofi Annan* has been in *England* since last Wednesday.
  - *Bush* is expected to hold on to *the White House* for the next four years.
- Finite verbs with habitual aspect
  - *Bill Clinton* and his family ski in *Aspen* regularly.
- Implicits
  - Implicit relations with time attributes
    - *Israeli policemen* fired live rounds in the air Thursday to disperse hundreds of young Palestinians who blocked *a major West Bank road* to show their support for Saddam Hussein.
  - Are there implicit time attributes?

Annotation Process

Annotation Staff

- Staff
  - Linguists, computational background helpful
- Training
  - Learn guidelines and tool
  - Training sets – comparisons
    - comparison viewer tool

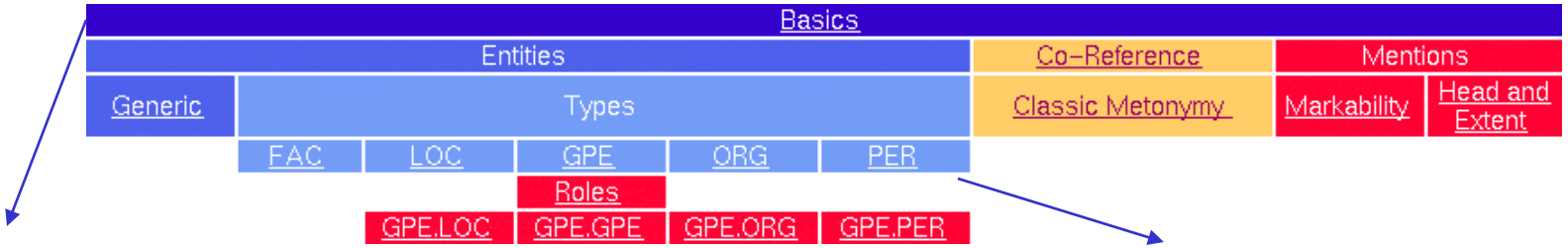


Project Management

- Discussion
  - Work environment
    - Annotators work side-by-side
    - Facilitates discussion of issues
    - Daily informal ACE chats
  - Project manager also involved in annotation
- Problem Log and Web Guidelines
  - Document annotator uncertainties

Documentation

- Integrated web-based EDT and RDC annotation guidelines
  - Part of annotator training
  - Plentiful examples with context taken directly from problem logs



Entity and Metonymy Basics

Fundamentals

For our purposes, an **entity** is some object in the world, and a **mention** is a reference to an object. An object may be referenced in several different ways. It can be called by its name, it can be indicated by a common noun or noun phrase (these are called "nominals"), or it can be represented by a pronoun. For example, the following are different mentions of an entity:

- **Name Mention:** Sean McGrew
- **Nominal Mention:** the guy who created this web page
- **Pronoun Mentions:** I, my

Our job is to find and categorize all the entities referred to by a document and link them to all their mentions.

In this project, we only care about certain kinds of entities: People, Locations, Facilities, Geo-Political Entities (GPEs), and Organizations constitute salient entities. We do not mark mentions of animals or most inanimate objects.

Classification

Entities

There are two decisions to make for each entity:

- What is the **entity type**: PER, LOC, FAC, GPE, or ORG?
- Is this entity **generic** or **specific?**

We assign each entity a type. A person or group of people is of type person (PER), unless the group has enough structure to qualify as an organization (ORG) or a Geo-Political Entity(GPE). A building or man-made structure is type facility (FAC). An astronomical body or definable area on the land or water is a location (LOC) unless it qualifies as a GPE. A GPE includes the government, land, and people of a definite region. GPEs ordinarily consist of a city, county, state, or nation, or a group of GPEs.

Base Type: Person (PER)

Definition

Person entities must be humans. They can be single individuals or a group if the group has a group identity.

People may be specified by **name, occupation, family relation, pronoun,** etc., or by some combination of these. **Dead people and human remains** are to be recorded as entities of type person. So are **fictional human characters** appearing in movies, TV, etc.

Mentions of **unborn children and/or fetuses** are a little tricky. These mentions will occur most often in articles about abortion rights or related topics. We do not want to get into that sort of a political discussion with this project. If the mention is in a quote, follow the meaning intended by the speaker. If the mention is not in a quote, but the article clearly leans one way or another, mark the mention according to the intended meaning. In general, fetuses are not marked as PER.

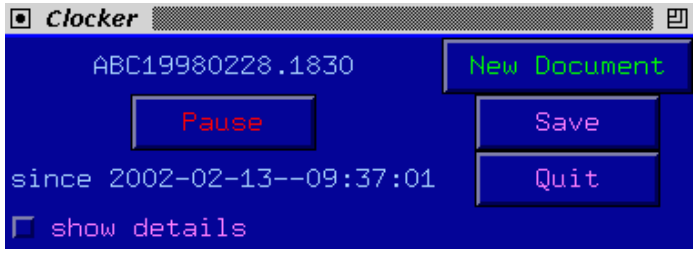
Groups of people are to be considered entities of type person unless the group meets the requirements of an ORG or GPE. For more information about this distinction, see below.

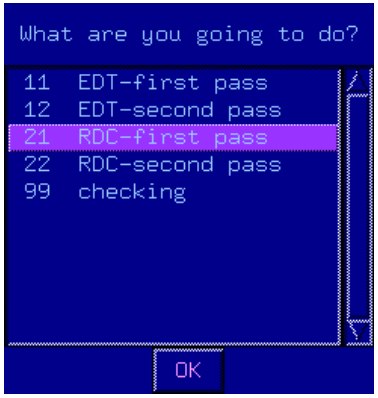
Examples

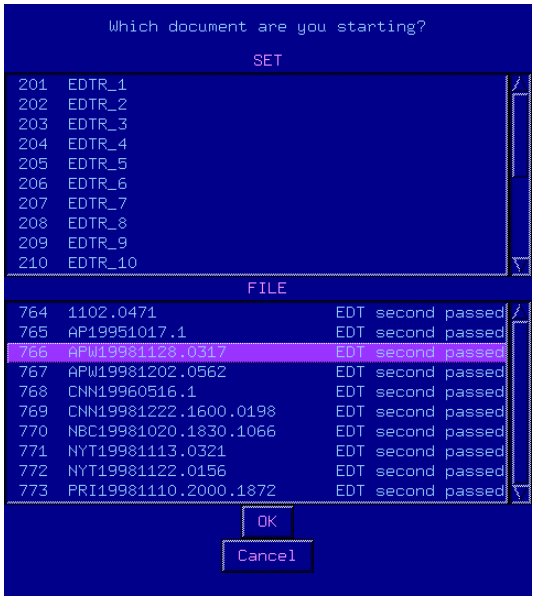
- **Valid PER Mentions**
  - *John Smith*
  - *the **butcher***
  - *dad*
  - *he*
  - *the **family***
  - *the house **painters***
  - *the **linguists** under the table in the living room*

Annotation Process

- EDT and RDC annotation with Alembic Workbench
- Two complete passes through the data
- 5-10% dual annotation built in
  - Comparison
  - Discussion
- Clocker
  - Tracks status of file sets (10 files per set)
  - Records annotation time in database







Annotation Effort

- EDT annotation effort
  - Phase 1 train/dev data annotated by LDC
    - 60K words
    - 3810 entities
    - 9618 mentions
  - Minimum of 8 decisions/mention (plus GPE role, metonymy decisions for subset of mentions)
  - Conservatively, 77K annotation decisions
  - Annotation rate – 10 wpm
- RDC annotation effort
  - Phase 2 RDC train/dev data
    - 210K words
    - 8151 relation pairs
  - Minimum 6 decisions/relation
  - Conservatively, 49K annotation decisions
  - Annotation rate – 30 wpm or better

Quality Assurance Measures

ACE Pilot Phase

- Initial set of working guidelines
- Annotation across multiple sites of small “test” sets
  - Triple annotation across 3 sites
  - Comparison and discrepancy resolution
  - To establish inter-annotator consistency and resolve guideline questions
- Refine guidelines based on annotation issues
- Ongoing communication via ace_list, email, conference calls
- Error reports from sites and adjudication/fixes

Local Annotation QC

- After Pilot Phase, additional QC measures adopted locally for both EDT and RDC

- Learning “test” sets
  - Dual annotation, discrepancy resolution for all files
    - Part of new annotator training

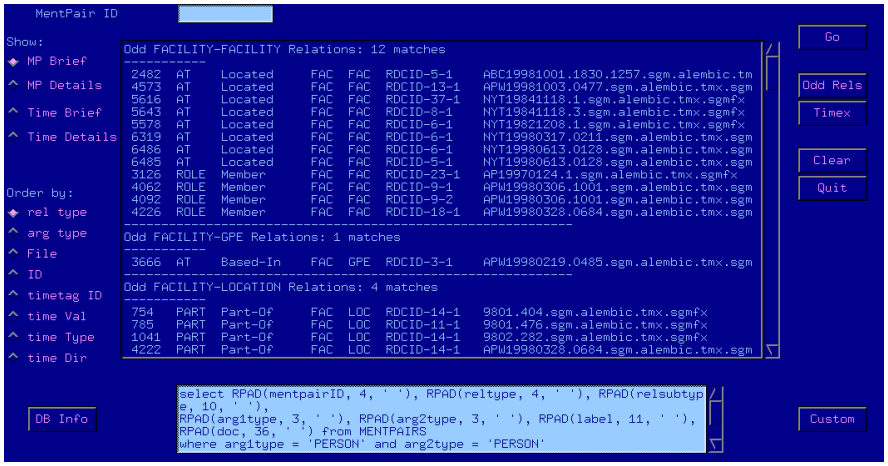
- Training, Development and Evaluation data
  - 5% dual annotation & discrepancy resolution
  - Second pass of all files
  - Additional “guided” third pass based on results of dual annotation discrepancies and problem logs
    - E.g., grep for keywords
    - Programmatic data scans

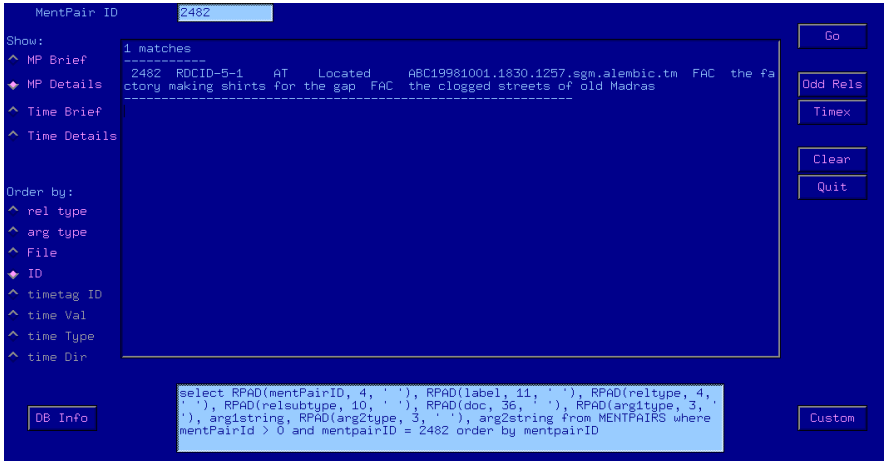
Annotation Errors

- Judgment calls/annotator world knowledge differences
  - Coref on multiple mentions of UN Inspectors
  - How many starting pitchers do the Atlanta Braves have?
- Clear annotation mistakes
  - Laredo, TX tagged with base type PER
  - [the eight sites, all of which are presidential compounds]
- Task-tool interface errors
  - Roles for PER base type – “Belfast’s ordinary folks”
  - Coreference in large files – “Israeli prime minister Netanyahu”, “the Israeli leader” spaced far apart
  - AT.Residence(FAC-GPE) – “The McDonald’s in the next town”
- Guideline ambiguities (mostly GPE, GPE-role)
  - Suburbs, regions, groups of GPEs – LOC or GPE?
  - Persian Gulf – LOC or GPE?

RDC QC: Tools

- Relquery
  - General and detailed views
  - Guide for second pass and programmatic scans
  - RDC only (EDT version in the works)
- Programmatic Scans
  - Type checking for relations
  - Data format
  - Common errors





Conclusions and Future Plans

LDC's evolving role in ACE

- Sole annotation site for future ACE evals
- Larger role in guidelines development and maintenance
  - Revise RDC guidelines to include EE-targeted relations
  - Revise EDT, metonymy, generics guidelines to provide unified annotation specification
  - LDC as "keeper of the guidelines"
    - Provide web guidelines created locally for annotator training
    - Incorporate updates, revisions as needed
- Create annotation task definition
  - Specifying annotation procedures, tools, QC, timeline

Future ACE plans (1)

- Ongoing modifications to EDT, RDC tasks
  - Rework GPEs, roles
  - EDT revisions
    - Expand entity types
      - E.g., Artifacts?
  - RDC revisions
    - Expand temporal attributes
  - Event detection and characterization
  - Continued collaboration with Evidence Extraction community

Future ACE plans (2)

- Add new languages
  - Chinese, Arabic in next phase of ACE
    - Requires guidelines modifications
    - Data availability
- Annotation tool development
  - LDC programmers developing Annotation Graphs-compliant tool
    - Multilingual
    - Data format
      - Current tool requires conversions – leads to problems
    - Simple, clean and user-friendly
    - Automatic error checking
      - Data format, missing tags, etc.