# Background and Context for CLASP

Nancy Ide, Vassar College

# The Situation

- Standards efforts have been on-going for over 20 years
- Interest and activity mainly in Europe in 90's and early 2000's
  - Text Encoding Initiative (TEI) – 1987
    - Still ongoing, used mainly by humanities
  - EAGLES/ISLE
    - Developed standards for morpho-syntax, syntax, sub-categorization, etc. (links on CLASP wiki)
    - Corpus Encoding Standard (now XCES - http://www.xces.org)

# Main Aspects

- Harmonization of *formats* for linguistic data and annotations

- Harmonization of *descriptors* in linguistic annotation

- These two are often mixed, but need to deal with them separately (see CLASP wiki)

# Formats: The Past 20 Years

| | |
|---|---|
| 1987 | TEI |
| | |
| 1994 | MULTEXT, CES |
| | |
| ~1996 | XML |
| 2000 | ISO TC37 SC4 |
| 2001 | LAF model introduced |
| | |
| now | LAF/GrAF, ISO standards |

Myriad of formats

Myriad of formats

# Actually…

- Things are better now
  - XML use
  - Moves toward common models, especially in Europe
  - US community seeing the need for interoperability
  - Emergence of common processing platforms (GATE, UIMA) with underlying common models

# Resources

1990

- WordNet gains ground as a "standard" LR
- Penn Treebank, Wall Street Journal Corpus
- British National Corpus
- EuroWordNet
- Comlex
- FrameNet
- American National Corpus
- Global WordNet
- More FrameNets
- SUMO
- VerbNet
- PropBank, NomBank
- MASC

present

World Wide Web

XML

Semantic Web

# NLP software

1994
1995
1996
1998

2003

200?

- MULTEXT > LT tools, LT XML
- GATE (Sheffield)
- Alembic Workbench
- ATLAS (NIST)
  - What happened to this?
- Callisto
- UIMA

Now: GATE and UIMA widely used, interoperable

# Where are we now

- We’ve learned a lot from past experience
- Technologies are vastly changed
  - Web technologies
  - distributed data and processing
  - formal models (maybe)
- Need for standards within the international community more urgent as access increases

# Recent US Interest

- In the past few years the US community has become interested in (at least some levels of) standardization
- Motivations:
  - Need to create and merge annotations at different linguistic levels in order to study interactions and interleave processing
  - Need to develop data and tools for emerging and strategic languages such as Chinese and Arabic, and minor languages
  - Need to make a major leap in the productivity of NLP research and language processing capabilities

# Recent Major Activities

- **Formation of ISO TC37 SC4** to develop a linguistic annotation framework and standard representation formats for various types of linguistic annotation
- Global efforts to create **linked wordnets and framenets**
- Development and harmonization of **systems and frameworks for linguistic annotation** (e.g., GATE, UIMA)
- **Recent major meetings** devoted to resource interoperability
  - CyberLing (link on CLASP wiki) E-MELD, TILR
  - International conference devoted to language resource interoperability (ICGL)
  - Multiple workshops at major conferences addressing issues of standards for representation formats and linguistic categories

- Establishment of **registries and catalogues for linguistic categories** (e.g., ISO TC37 SC4 data category registry) and annotation schema (e.g., UIMA component registry)
- **U.S.-funded efforts** to merge and/or harmonize linguistic annotations at different levels (OntoNotes, Unified Linguistic Annotation), and different phenomena (WordNet and FrameNet)
- **EU-funded effort** to create a common resource and infrastructure for the humanities and social sciences (CLARIN)
- **Formation of an ACL special interest group** (SIGANN), with a primary aim to work toward the development of standards for representing and designating linguistic information
- Independent work within the Semantic Web community on interoperability of ontologies

# SILT

- Sustainable Interoperability for Language Technology
- Funded by National Science Foundation’s INTEROP program
- PIs: Nancy Ide, James Pustejovsky
- Parallel EU project: FLaReNet
- Efforts to involve Asians

- http://www.anc.org/SILT

# SILT Goals

- Survey of resources, tools, and frameworks_
  - Examine what exists and what needs to be developed
  - Identify areas for which interoperability would have the broadest impact in advancing R&D
- Identify major standards/interoperability efforts and existing and developing technologies
  - Examine ways to leverage results to define an interoperablity infrastructure for tools and data
- Analyze innovative methods and techniques for the creation and maintenance of language resources in order to
  - Reduce high costs
  - Increase productivity
  - Enable rapid development of resources for new languages

# SILT Goals

- Implement proposed standards and best practices in corpora currently under development (e.g., American National Corpus, TimeBank)
  - Evaluate their viability
  - Feed into the process of standards development
  - Test and use interoperability frameworks (e.g. UIMA), and implement processing modules
  - Distribute all software, data, and annotations

# ISO effort

- International Standards Organization (ISO) sub-committee on Language Resource Management (ISO TC37 SC4)
- Goal: define standards for representing linguistic annotations and other resources
  - incorporate *de facto* standards and "best practices" into a coherent whole

# ISO TC37 SC4 Working Groups

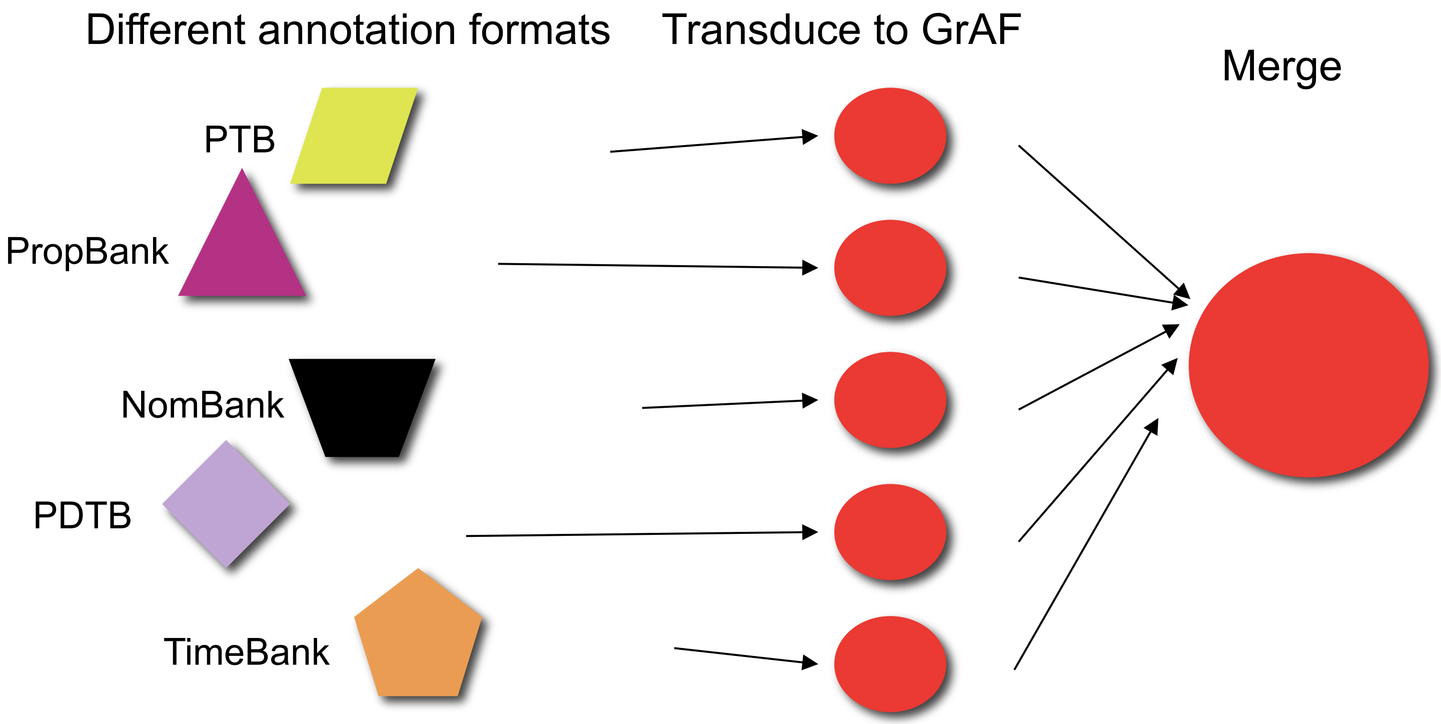
- Linguistic Annotation Framework (Nancy Ide)
  - Underpinning of all standards in SC4 for format and architecture
- Morphosyntactic Annotation Format
- Syntactic Annotation Format
- Word Segmentation
  - Only Asian languages at present
- Semantic Annotation
  - Time and Events (James Pustejovsky)
  - Semantic Roles (Martha Palmer)
  - Space (James Pustejovsky)
- Feature Structures

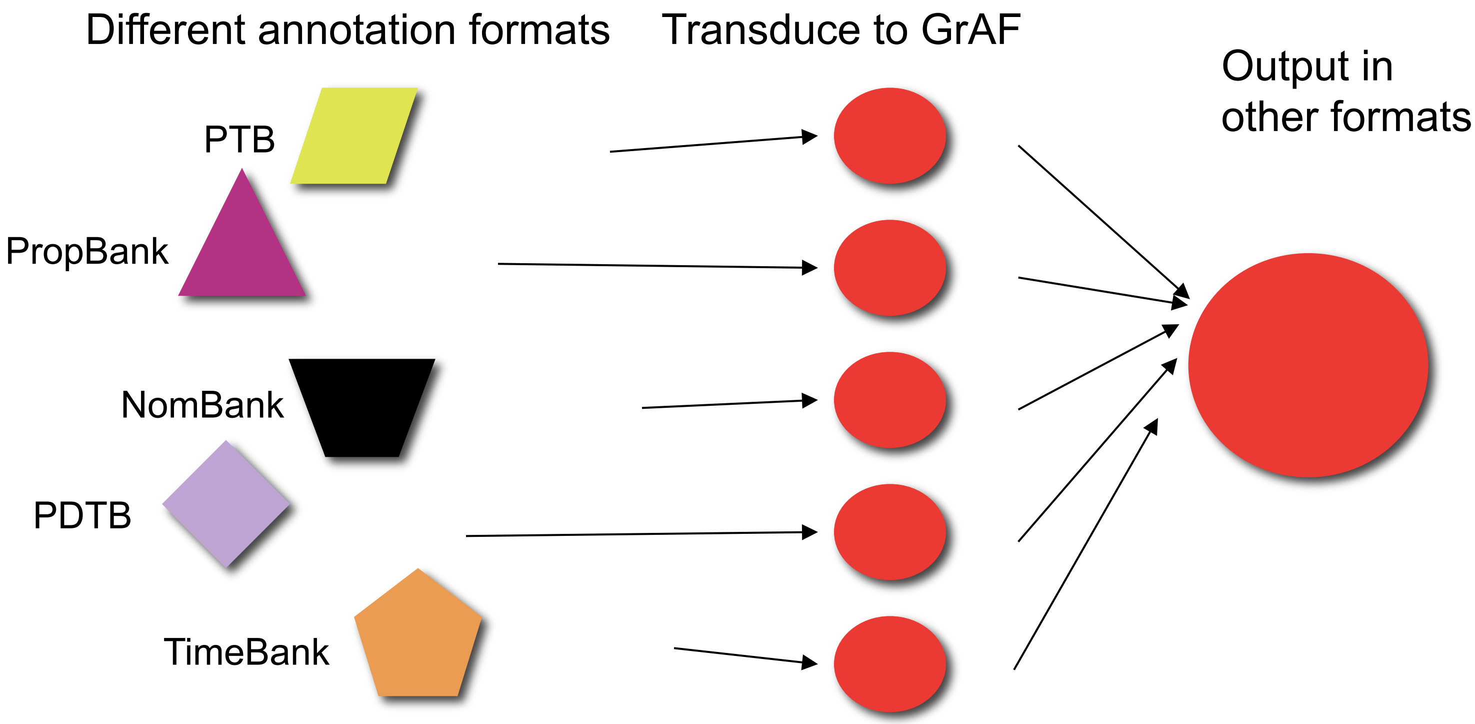
# Linguistic Annotation Framework

- Provides a "pivot" format for annotations
- Map existing formats into the pivot
- Pivot: XML serialization of a graph decorated with feature structures

- MASC is an implementation:
  - Multiple annotations contributed from diverse sources
    - Penn Treebank, FrameNet, GATE's noun and verb chunkers and named entities, PropBank (soon: TimeML, BBN named entities, HPSG, Penn Discourse Treebank, and others)
    - All transduced to LAF (GrAF) format
    - Can be merged, output in other formats if desired

    NB: alternative tokenizations have plagued us! We hope to avoid aligning tokenizations in the future…


Different annotation formats
Transduce to GrAF
Merge
PTB
PropBank
NomBank
PDTB
TimeBank



Different annotation formats

Transduce to GrAF

Output in other formats

PTB

PropBank

NomBank

PDTB

TimeBank

XCES (XML inline), XIARA, word/POS, UIMA, NLTK…

# ISOCat

- The ISO Data Category Registry
- Addresses issue of standardization of annotation content
- Provides a set of reference categories onto which scheme-specific names can be mapped
- Provides a precise semantics for annotation categories
- Provides a point of departure for definition of variant, more precise, or new data categories

# Exchange Specification

- Annotations may use ISOCat categories directly (via PID) or provide a mapping between scheme-specific instantiations and concepts in the Data Category Registry
  - Document departures, variations, additions
- Used in data exchange
  - provides receiver with information to interpret annotation content or map to another instantiation
  - semantic integrity guaranteed by mutual reference to DCR concepts or definition of new categories by annotator

# Annotation Layers

- Conceptual layers of annotation
  - E.g. morpho-syntax, syntax, co-reference…
  - SC4 defining a set of layers
- Each layer has a schema defining the relevant categories and relations
  - E.g. syntax
    - Category: Sentence
    - Relations: SUBJ (Object: NP), MainVerb (Object:VP), “Constituent” (Object: NP |VP | PP)
- Inter-layer and cross-layer relations

# Goals

- Reference categories in ISOCat rather than give cats
- Reference FS fragments and schema layer definitions in on-line libraries

# Comments for CLASP

- Our focus is primarily on linguistic descriptors (categories)
  - Is the ISOCat model (or ISOCat itself) the way to go?
  - Would the US community buy in to this sort of approach?

# Segmentation (tokenization)

- Some de facto standards for formats are emerging that affect decisions about tokenization
  - Stand-off annotation
    - No need (in fact, prohibition) to segment in-line (change data)
    - Tokenization considered an annotation
    - Can have multiple tokenizations of same data
    - Can skip issues of where to break words etc. such as "can't" by simply associating (via links) two tokens (e.g. "can" and "not") with the string

- LAF approach to segmentation
  - Segmentation is an annotation
  - Data is "read-only": corrections, normalizations, etc. all treated as annotations

- Recommendation: Tokenization standards developed as a part of/contributed to ISO working group on word segmentation

# Cannot afford to be "US-centric"

- Standards cannot be developed in isolation of what has been done and is being done in the rest of the world
  - E.g., Penn Treebank tokenization and POS is far from a universal anywhere else
  - Must develop standards with an eye toward their use in other languages so that we allow for the potential to combine multi-lingual data
    - Tokenization rules for English won't necessarily work for other languages, or even generalize
  - Take into account the vast amount of work already done elsewhere so as not to reinvent the wheel (again)