Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York USA

THE MANUALLY ANNOTATED SUB-CORPUS

An Experiment in Collaborative Language Resource Development

The Need

- Annotated corpora are a fundamental resource for research and development in the field of natural language processing (NLP)
- Need for corpora annotated for multiple phenomena across a variety of linguistic layers keenly recognized in the computational linguistics community
  - Linguistic annotations provide a richer set of features for machine learning than unannotated corpora (Gigaword, Wikipedia, etc.) -- potentially better language models
  - Deeper linguistic information and study of intra-level interactions may lead to insights that can improve NLP applications

The Need

- Need for corpora spanning multiple genres and domains, including new genres (tweets, blogs, email…)
- Need for annotated corpora that are both easily accessible and available for use by anyone

- However, contemporary, multi-genre language corpora with high-quality annotations for diverse linguistic phenomena that are openly available are relatively rare
    - Even for English!

Existing Corpora of English

- Brown Corpus
  - Freely available (must request and sign license)
  - Broad genre
  - Only part of speech annotation
  - Some added annotations in diverse formats
    - Not usable together
  - Small: 1 million words
  - “Old” language (mid-1960’s)

Existing Corpora of English

- British National Corpus
  - Most widely used
  - Freely available (must request and sign license)
  - Broad genre
  - Large: 100 million words
  - British English only
  - Only part-of-speech annotation
  - Not contemporary (up to ~1990)
    - No modern genres
    - The word “browser” not included!

Existing Corpora of English

- Corpus of Contemporary American English (COCA)
  - Available via web interface
  - Huge! 450 million words
  - Web data
  - Only part of speech available (done on the fly)
  - Can only access concordances of individual sentences due to copyright restrictions on web data
    - Cannot use for training language models
    - Cannot add annotations
    - Cannot study discourse-level phenomena

Existing Corpora of English

- Penn Treebank
  - Most well-known, one million word *Wall Street Journal* corpus
  - Over the years, fully or partially annotated for several phenomena, but in a variety of different formats
    - Difficult to combine annotations in order to study inter-relations
  - Not free; license from LDC
  - Limited genre long recognized as a problem
    - Highly stylized text, limited syntactic variation
    - Only one sense of the word “stock”!

Existing Corpora

- OntoNotes (English portion)
  - License from LDC, no cost
  - Multiple annotations in a common format
    - Penn Treebank syntax, PropBank predicate argument structures, co-reference, named entities
  - Small: 1 million words
  - Limited genres: newswire, broadcast news, broadcast conversation
  - Limited to annotations produced by members of the OntoNotes project
    - Annotations cannot be added by others
  - Use of the data and annotations with software other than the OntoNotes database API not straightforward

Existing Corpora of English

- Open American National Corpus
  - Freely downloadable from web
  - Broad genre
  - Contemporary language (1990-present)
  - Medium size (15 million words)
  - Several layers of annotations in a common format
  - Can add annotations
    - Contributed annotations rendered into same common format
  - Annotations largely unvalidated

The Problem

- High quality corpora and annotations costly to produce
  - Steps required:
    - Acquire appropriate and available language
    - Prepare data originally in a variety of formats
    - Remove formatting, interspersed HTML, etc.
    - Perform manual annotation or manual validation for automatically produced annotations
      - Provide environment for accomplishing manual work
      - Ideal to use multiple annotators under controlled circumstances to provide inter-annotator agreement measures, esp for semantic/discourse level annotations

Bottom Line

- Corpus development can require several man-years of labor-intensive effort and substantial funding
  - Substantial funding for resource development is difficult to acquire
  - Production and annotation of corpora not always a recognized scientific activity, researchers hesitant to undertake the task

The Solution?

- Distribute effort among members of the research community
  - Distribute cost as well
- *Wall Street Journal* corpus annotation showed community willingness to undertake a distributed effort
  - But lack of coordination led to annotations that could not be used together

MASC

- US National Science Foundation (2007-2012) project to produce a Manually Annotated Sub-Corpus of the Open American National Corpus

- Goals:
  - Offset high costs of producing high quality linguistic annotations via a distribution of effort
  - Solve usability problems for annotations produced at different sites by harmonizing their representation formats
  - Ensure all data and annotations freely and easily obtained for any use

MASC

- Much wider variety of genres than existing multiply-annotated corpora of English
- All data drawn from contemporary American English (1990-)
- Fully open model of distribution, without restriction, for all data and annotations
- MASC project committed to incorporating diverse annotations contributed by the community, regardless of format

MASC is the first large-scale, open, community-based effort to create a much-needed language resource for NLP

The Corpus

- 500K words of contemporary (1990 on) American English
- Completely open data and annotations
- Nineteen genres
  - Evenly balanced across genres
  - 15% spoken transcripts, 85% written
- Sixteen types of annotation
  - 10 types on the entire MASC
  - 3 additional types on 40-55K words
  - 3 additional types on ~5K words

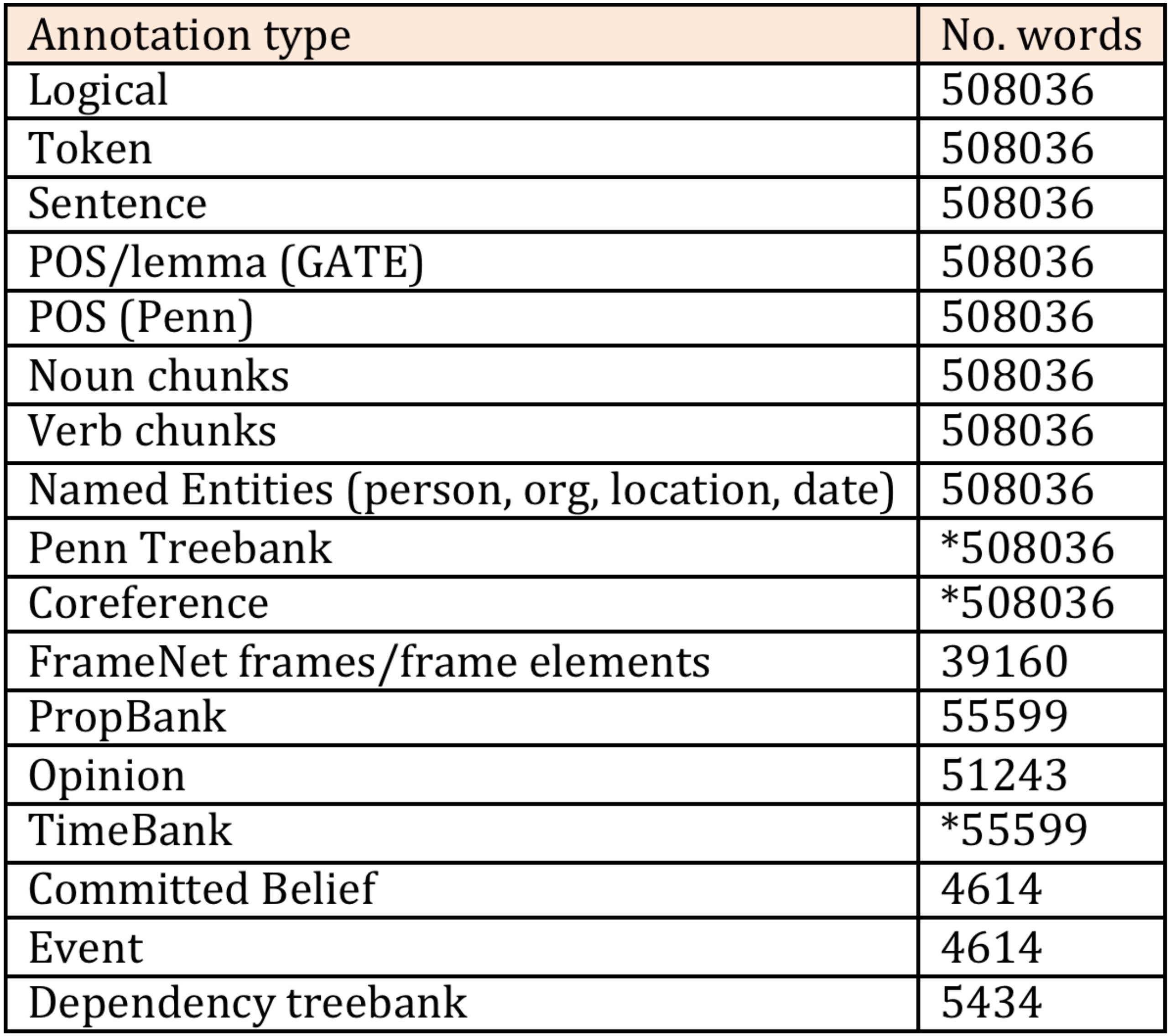
MASC Genres

Genre	No. files	No. words	Pct corpus
Court transcript	2	30052	6%
Debate transcript	2	32325	6%
Email	78	27642	6%
Essay	7	25590	5%
Fiction	5	32811	7%
Gov't documents	5	24578	5%
Journal	10	25635	5%
Letters	40	23325	5%
Newspaper/newswire	41	23545	5%
Non-fiction	4	25182	5%
Spoken	11	25783	5%
Technical	8	27895	6%
Travel guides	7	26708	5%
Twitter	2	24180	5%
Blog	21	28199	6%
ficlets	5	26299	5%
movie script	2	28240	6%
spam	97	23465	5%
jokes	16	26582	5%
TOTAL	363	508036	

Annotations

- Manual annotations or manually-validated annotations on whole corpus for multiple levels
  - Word and sentence boundaries, part of speech (three different versions)
  - Shallow parses (noun and verb chunks)
  - Named entities
    - Enables linking WordNet senses and FrameNet frames into more complex semantic structures
    - Enriches semantic and pragmatic information
  - Penn Treebank syntax
  - Coreference
- Other annotations on parts of corpus: full text FrameNet annotation, PropBank, NomBank, Opinion, etc.

MASC Annotations



Annotation type	No. words
Logical	508036
Token	508036
Sentence	508036
POS/lemma (GATE)	508036
POS (Penn)	508036
Noun chunks	508036
Verb chunks	508036
Named Entities (person, org, location, date)	508036
Penn Treebank	*508036
Coreference	*508036
FrameNet frames/frame elements	39160
PropBank	55599
Opinion	51243
TimeBank	*55599
Committed Belief	4614
Event	4614
Dependency treebank	5434

** In progress*

Annotation Process

- Smaller portions of the sub-corpus manually annotated for specific phenomena
  - Maintain representativeness
  - At least two annotators do same data
- Apply (semi)-automatic techniques to determine reliability of results
- Study inter-annotator agreement on manually-produced annotations
  - Determine benchmark of accuracy
  - Fine-tune annotator guidelines

Bootstrapping automatic annotation

- Apply iterative process to maximize performance of automatic taggers
  - Identify common errors during manual annotation
  - Modify automatic annotation software to fix
  - Regenerate annotations
  - Semi-automatically evaluate results

- Improved annotation software later applied to the entire OANC
  - Provide more accurate automatically-produced annotation of full corpus

Annotation interactions

- Can accurate annotations for one phenomenon improve performance of automatic annotation systems for another?
    - E.G., Validated WN sense tags and noun chunks may improve automatic semantic role labeling

Alignment of Lexical Resources

- Concurrent NSF-funded project investigating how and to what extent WordNet and FrameNet can be aligned
- MASC annotations of FrameNet frames and frame elements and WordNet senses provide a ready-made testing ground

MASC Sentence Corpus

- Accompanying corpus of 1000 sentences for each of 114 words manually annotated for WordNet senses
- 100 sentences for each word annotated for FrameNet frames
  - Basis for WN-FN harmonization effort
- Uses “enhanced” WordNet sense inventory
  - Version 3.1 and beyond
  - MASC sense annotation task used to improve WordNet
    - Annotator feedback used to modify WordNet sense inventory

MASC Sentence Corpus

- Differences from existing sense corpora
  - Includes a large number of instances of each word
    - More than a few examples of less frequent senses
    - Includes all occurrences of each word in MASC, filled out to 1000 (where necessary) with sentences from OANC
  - Includes moderately polysemous words (~ 5 to 20 senses, average ~7)
  - Roughly balanced for POS
    - 30 Adjectives, 40 Nouns, 44 Verbs

- Distributed with full inter-annotator agreement data

MASC Format

- The layering of annotations over MASC data dictates the use of a stand-off annotation representation format
    - Each annotation contained in a separate document linked to the primary data
- All annotations represented using the ISO Graph Annotation Format (GrAF)
    - XML serialization of abstract model : directed graph decorated with feature structures providing the annotation content
    - Enables merging annotations originally represented in different formats
    - Enables generating annotations in a variety of other formats
    - Enables concurrent annotations of the same type

Linguistic Annotation Framework

- Developed in ISO TC37 SC4
- Now ISO standard: ISO 24612
- In addition to representation format, includes specifications for a Data Category Registry (ISOCat)
  - Repository of common linguistic categories for reference from annotations

Principles

- Separation of data and annotations
  - Stand-off annotation
- Separation of user annotation formats and the exchange (“dump” or pivot) format
  - Mappable to one another
- Separation of annotation structure (relationships among parts) and content (data categories) in representation of annotations
- Separation of user format and pivot

Abstract Data Model

- A standard where users can use any format they choose?
- Quid pro quo
  - User-defined formats must conform to an abstract data model (ADM) for annotations
  - I.e., must be mappable to the ADM

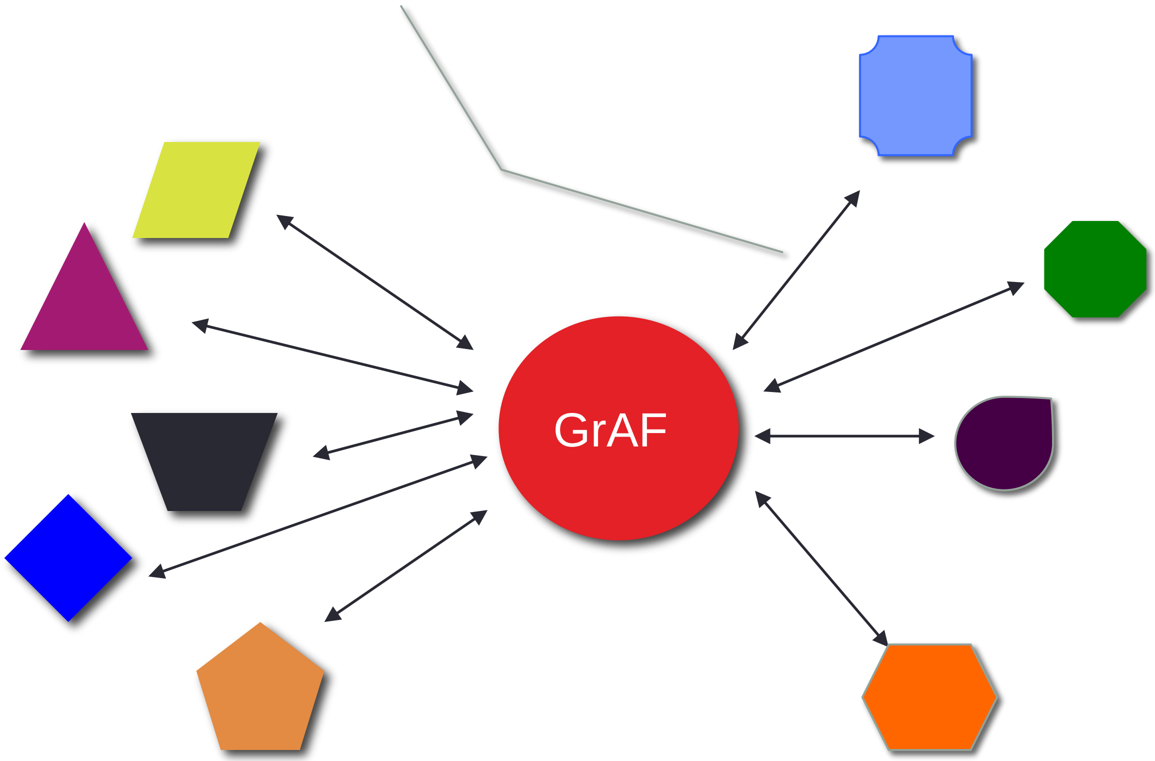
ADM

- Annotations represented as a graph of feature structures
  - Directed graph referencing *n*-dimensional regions of primary data and (possibly) other annotations
  - Nodes are labeled with feature structures containing the annotation content

Pivot Format

- ADM instantiated in a pivot format in XML (GrAF)
- Annotations are mapped to pivot for the purposes of exchange
- Pivot format version then can be transduced to other formats
- Each use format need only be mapped into and out of pivot to enable transduction to any other format

Converter into GrAF and out of GrAF provides interoperability with all other formats



Primary Data

- Primary data contains no annotations
  - “Read-only”
  - Modifications can be regarded as annotations
- Insistence on the identification of a base segmentation of the primary data
  - Identifies contiguous sequences of indivisible logical units
    - For text, usually a character
  - “Compatible” annotations (i.e. those that can be merged etc.) use common base segmentation

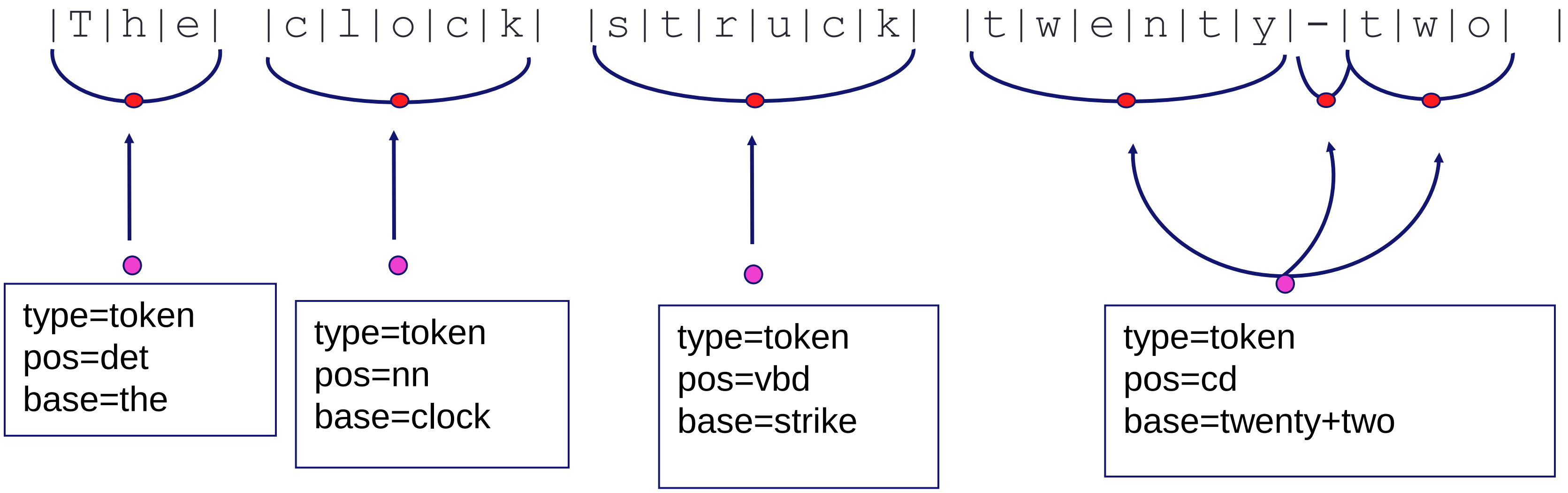
Segmentation

- Set of disjoint edges over primary data
    - Vertices
        - Virtual, located between each logical unit
        - Sequentially numbered
    - Edges
        - Each edge (x,y) in the graph delimits a non-divisible region of primary data
            - Can be beginning-end of text span, or several points in image, video, etc.

Segmentation

- Multiple segmentations may be defined over a single primary data set
  - Specify segmentations at different levels of granularity
  - A segmentation is “primary” vis a vis a given annotation, not the data itself
- Edges in a primary segmentation can be defined over any region of contiguous primary data, regardless of its length
- No need for region to be contiguous
- For text, most common primary segmentation is the token

Edge graph over primary data



|T|h|e| |c|l|o|c|k| |s|t|r|u|c|k| |t|w|e|n|t|y|-|t|w|o| |

type=token
pos=det
base=the

type=token
pos=nn
base=clock

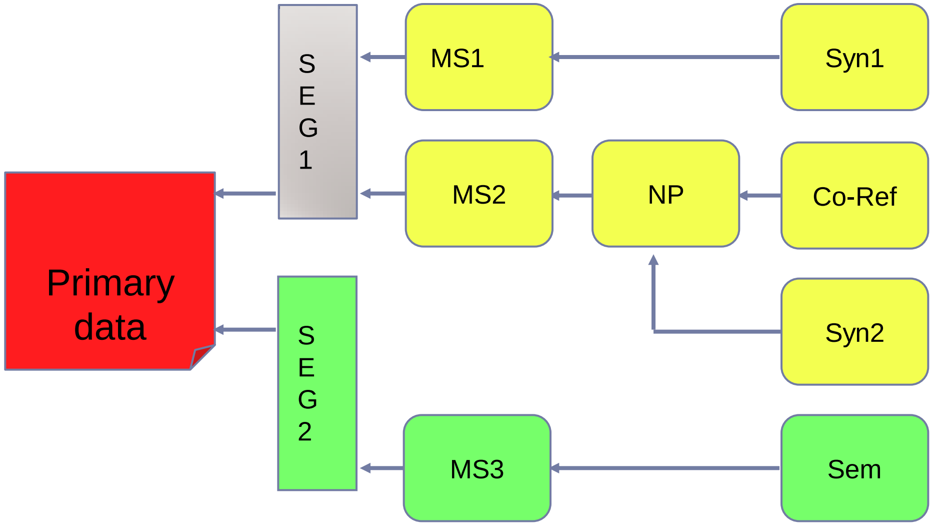
type=token
pos=vbd
base=strike

type=token
pos=cd
base=twenty+two

Annotation Layers

- Each annotation layer in a separate document that associates the elements in its content with a unique namespace
- Each annotation layer has a schema defining the relevant categories and relations
  - Map to type specification provided in annotation document header
    - May also define inter-layer relations

As many annotations as desired can reference the same segmentation or be layered over lower-level annotations



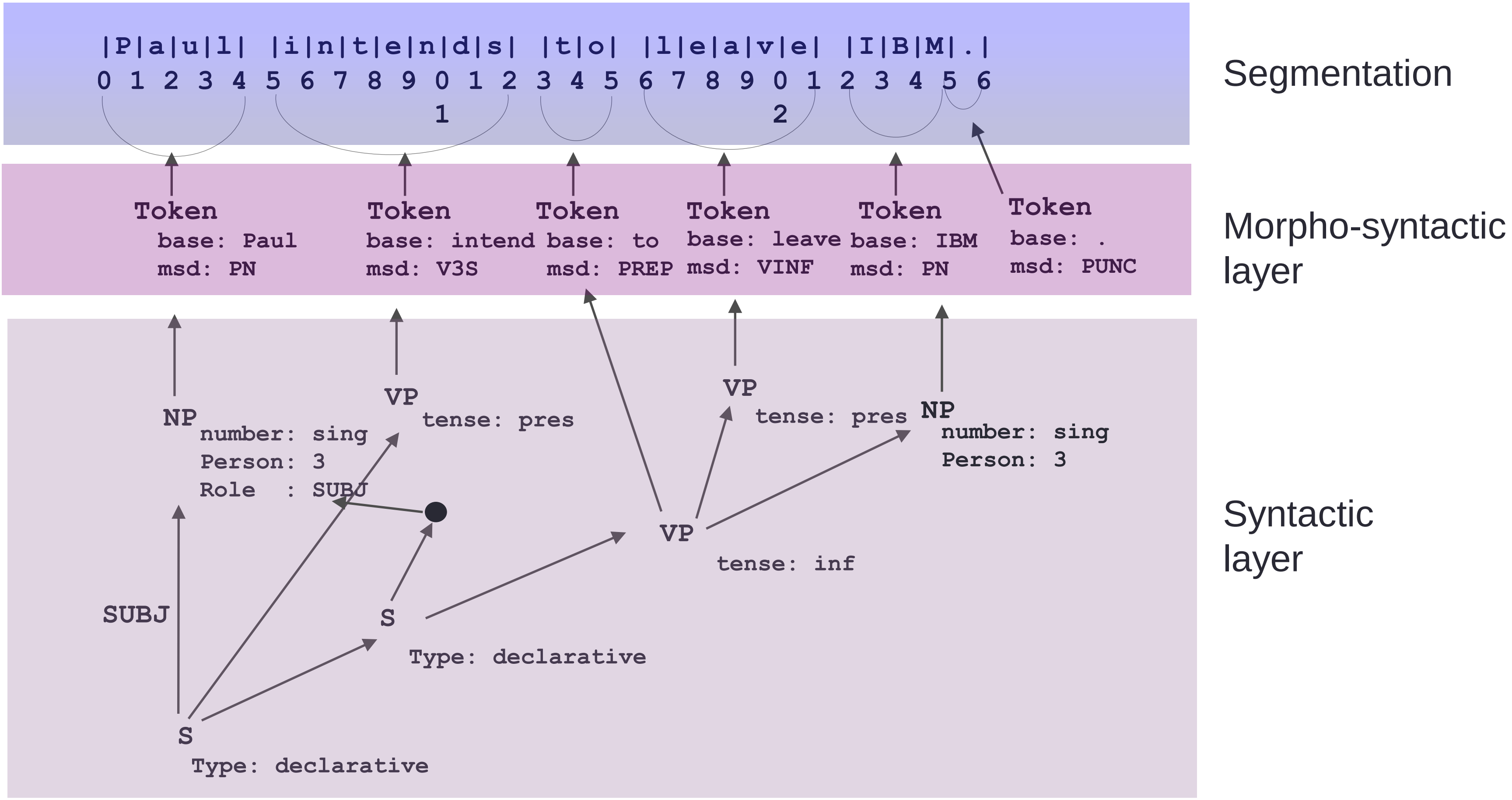
Feature Structures

- Annotation content associated with nodes in the graph represented as feature structures
- Encoded using ISO/TEI feature structure XML representation
  - A simplified version used for most cases
- Note: Feature structures are also graphs
  - Sub-graphs attached to nodes

Annotating Annotations

- Vertices in an annotation may be referenced from other annotations
    - Unlike annotation graphs
- The strategy described above may be applied recursively, thus creating a DAG whose leaves are the vertices of the segmentation

Annotation Layers



|P|a|u|l| |i|n|t|e|n|d|s| |t|o| |l|e|a|v|e| |I|B|M|.|

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6

1 2

Segmentation

Token base: Paul msd: PN

Token base: intend msd: V3S

Token base: to msd: PREP

Token base: leave msd: VINF

Token base: IBM msd: PN

Token base: . msd: PUNC

Morpho-syntactic layer

NP number: sing Person: 3 Role : SUBJ

VP tense: pres

VP tense: pres

NP number: sing Person: 3

VP tense: inf

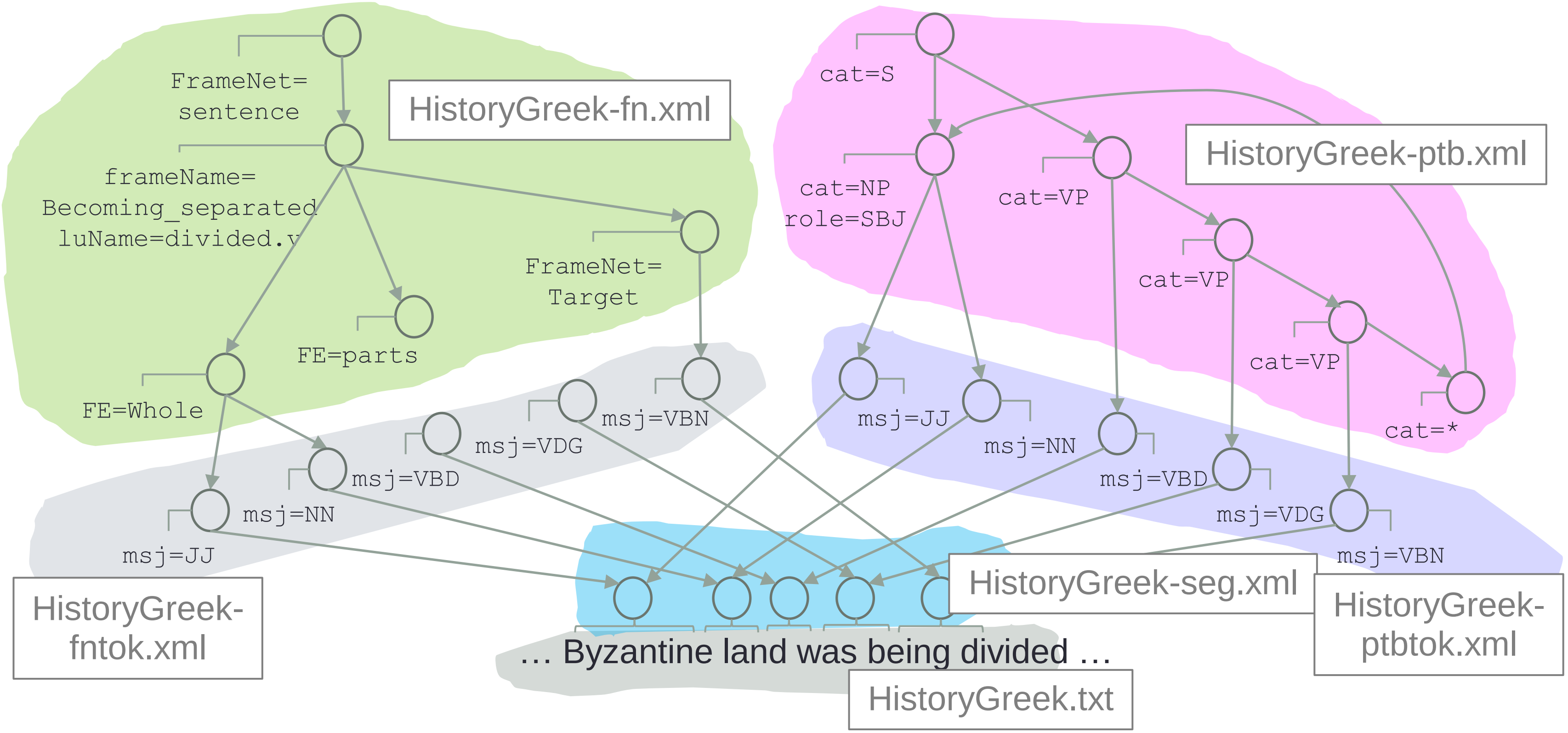
SUBJ

S Type: declarative

S Type: declarative

Syntactic layer

MASC in GrAF
e.g., frames and syntax



○ node (segment)

→ edge (relation, hierarchical)

↷ edge (relation, non-hierarchical)

┐ label attached to a node or an edge (annotation)

filename GrAF file (layer id)

Advantages of Graph Model

- Isomorphic to formats used by emerging annotation frameworks and tools
  - E.g., UIMA's Common Analysis System
  - Penn Discourse Treebank API
- Underlies Web formats such as RDF and OWL
  - Annotation graph is trivially transducable to their serializations (including XML and several others)
- Provides a well-understood model and basis for devising linguistic annotation schemes

Other Advantages

- Graph format makes it easy to
  - Add information
  - Modify graph to reflect additional analysis, correct errors, etc.
    - E.g., delete or move constituents such as punctuation and parenthetical phrases, conjoin sub-graphs joined by “and”, correct PP attachments based on information in the tree, etc.
  - Align in-line annotations of the same data
    - E.g., TimeBank’s version of the WSJ and the PTB’s version

ANC2Go

- Web application provided by the ANC project
  - Implemented as a RESTful web service
- Users choose data and annotations and desired output format
- Send request via web interface
- ANC2Go returns a URL from which the user can download the requested corpus
- Users create a “personalized” corpus
  - Data and annotations of their choosing
  - Format most useful to them

Output formats

- XML in-line
  - Suitable for use with the BNC’s XAIRA search and access interface
  - Input to any XML-aware software
- Token with part of speech tags, separated by character of the user’s choice
  - Input to general-purpose concordance software including MonoConc and Word- Smith
- Token/part of speech input for the Natural Language Toolkit (NLTK)
- CONLL IOB format
  - Used in the Conference on Natural Language Learning shared tasks
- Resource Description Format (RDF)
  - Basis of semantic web representations

Additional Formats

- Modules to use GrAF annotations in general-purpose annotation and analysis tools
  - GATE (plugins to read/write GrAF)
  - UIMA (CAS Handlers to read/write GrAF)
  - NLTK (corpus reader)
- Java API for GrAF (soon also Python API)
  - Use GrAF annotations directly
  - Includes a “GraphVizRenderer” to generate visualizations of an annotation subgraph
- Can use these systems independently or interchangeably
  - See Ide & Suderman, 2009, “Bridging the Gaps”

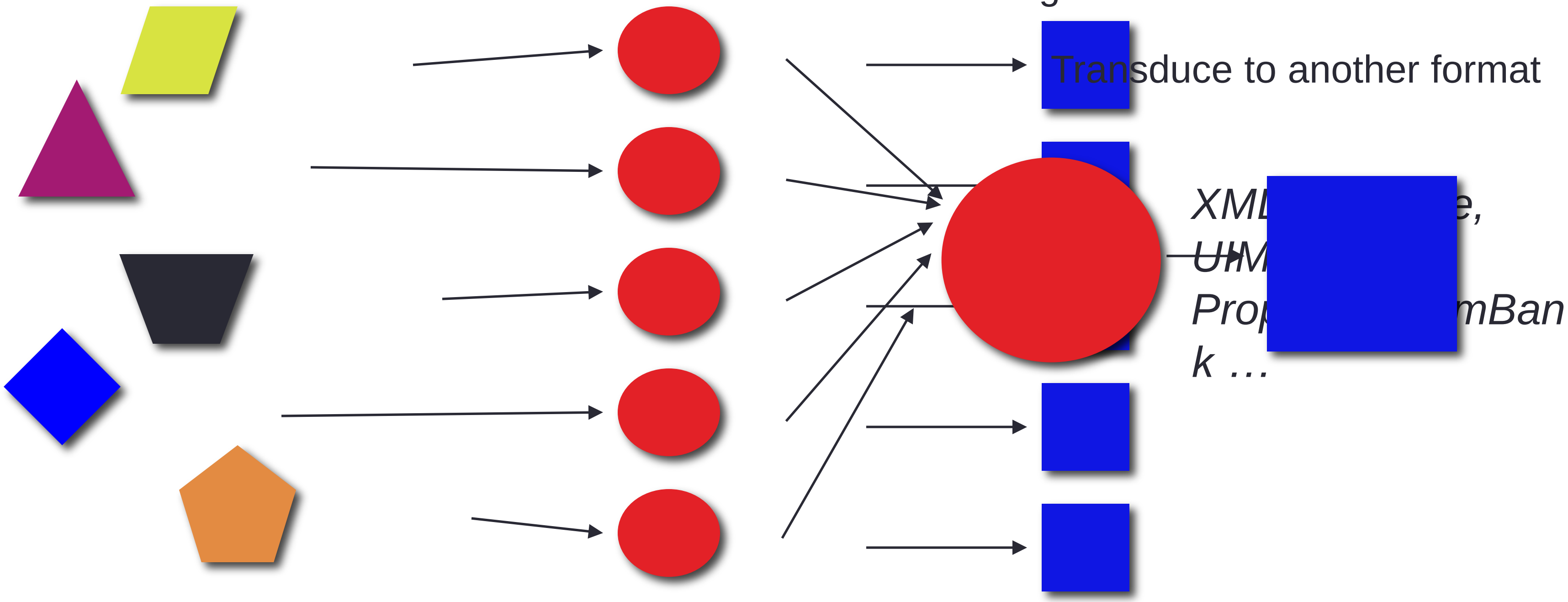
Transduction

Different annotation formats

Transduce to GrAF

Transduce to another format

Merge



Transduce to another format

XML

e,

UIM

Prop

mBan

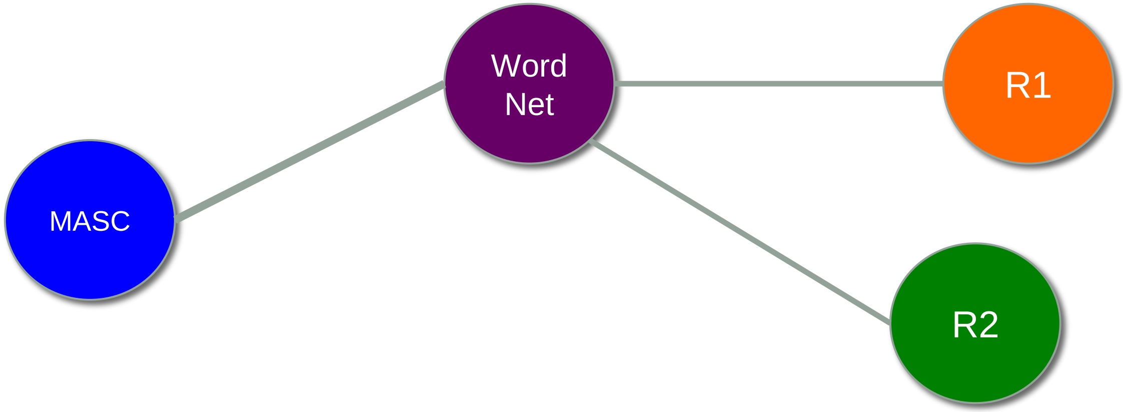
k …

MASC in the LLOD cloud

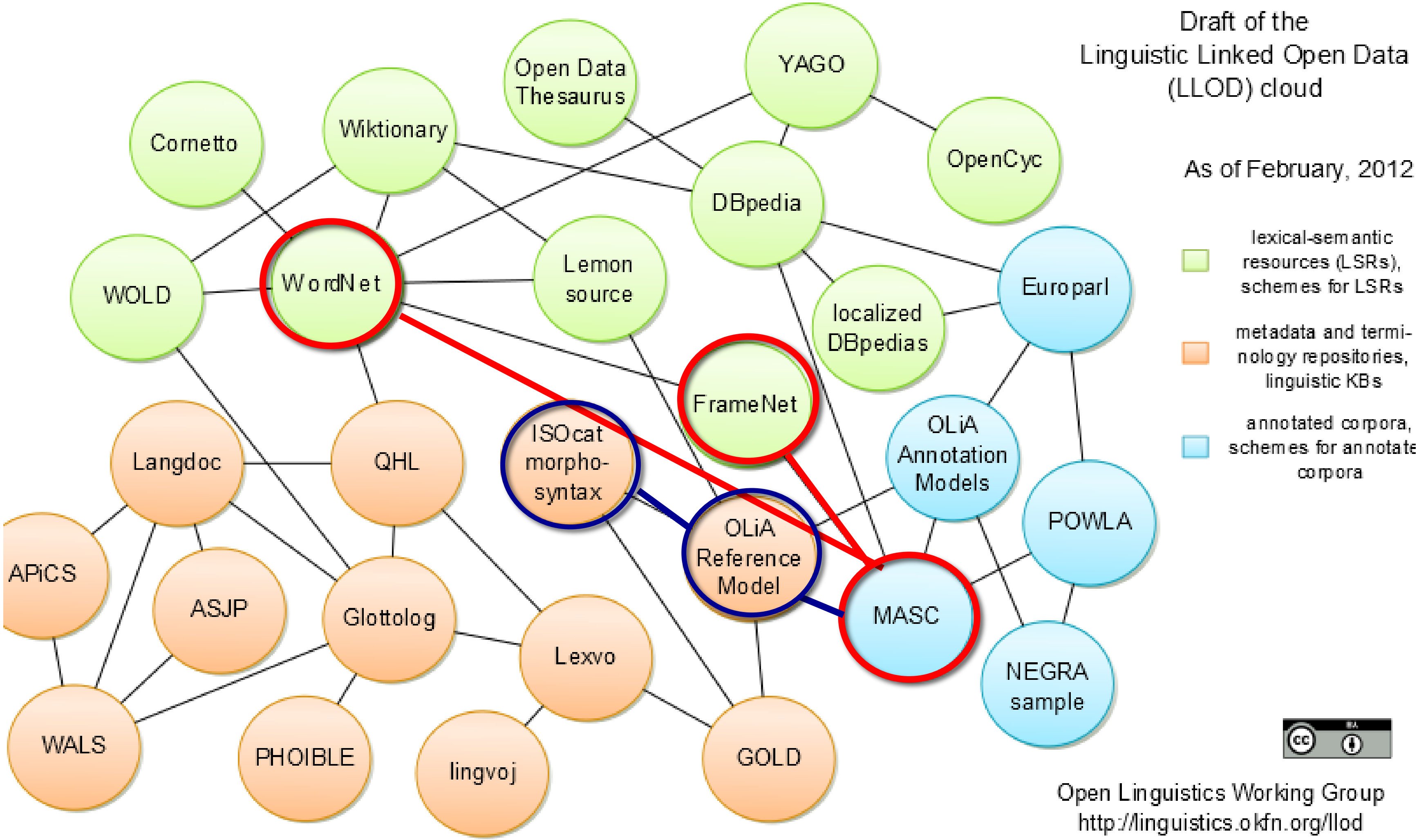
- Linguistic Linked Open Data (LLOD) cloud incorporates annotated language data and resources into the Semantic Web
  - Effectively, links all information across the web and provides for query access
  - Resources represented in RDF/OWL
    - Labeled links add semantics to web linkage
    - Isomorphic to GrAF; RDF link labeling technology is more concise
  - Enables exploitation of RDF/OWL tools (query) and other technologies

MASC in the LLOD cloud

- MASC is being transduced to RDF to be included in the LLOD cloud
  - Ideal because open (LLOD : “O” is for “open”)
  - Links from its annotations to their categories in WordNet and FrameNet, which have already been rendered into RDF/OWL
  - Via transitivity, links from MASC to WordNet and FrameNet also link it to other linked resources



MASC in the LLOD cloud



Draft of the
Linguistic Linked Open Data
(LLOD) cloud

As of February, 2012

🟩 lexical-semantic resources (LSRs), schemes for LSRs

🟧 metadata and terminology repositories, linguistic KBs

🟦 annotated corpora, schemes for annotate corpora

Open Linguistics Working Group
http://linguistics.okfn.org/llod

What can this do for NLP?

- Via links to WordNet and FrameNet can do RDF queries over Semantic Web such as

  “Find all tokens that refer to *land* as a political unit (synonyms from the WordNet synset *land%1:15:02::*) and fill the CONTENT slot in the FrameNet frame EXPERIENCER_FOCUS”

  (all places expressing an attitude towards one’s country)
- Generate (training) data for NLP tasks (wsd, semantic role labeling, sentiment analysis, etc.)
- Multiple views:
  - Many different annotations of same type can be compared/combined for better results
  - Different theoretical perspectives

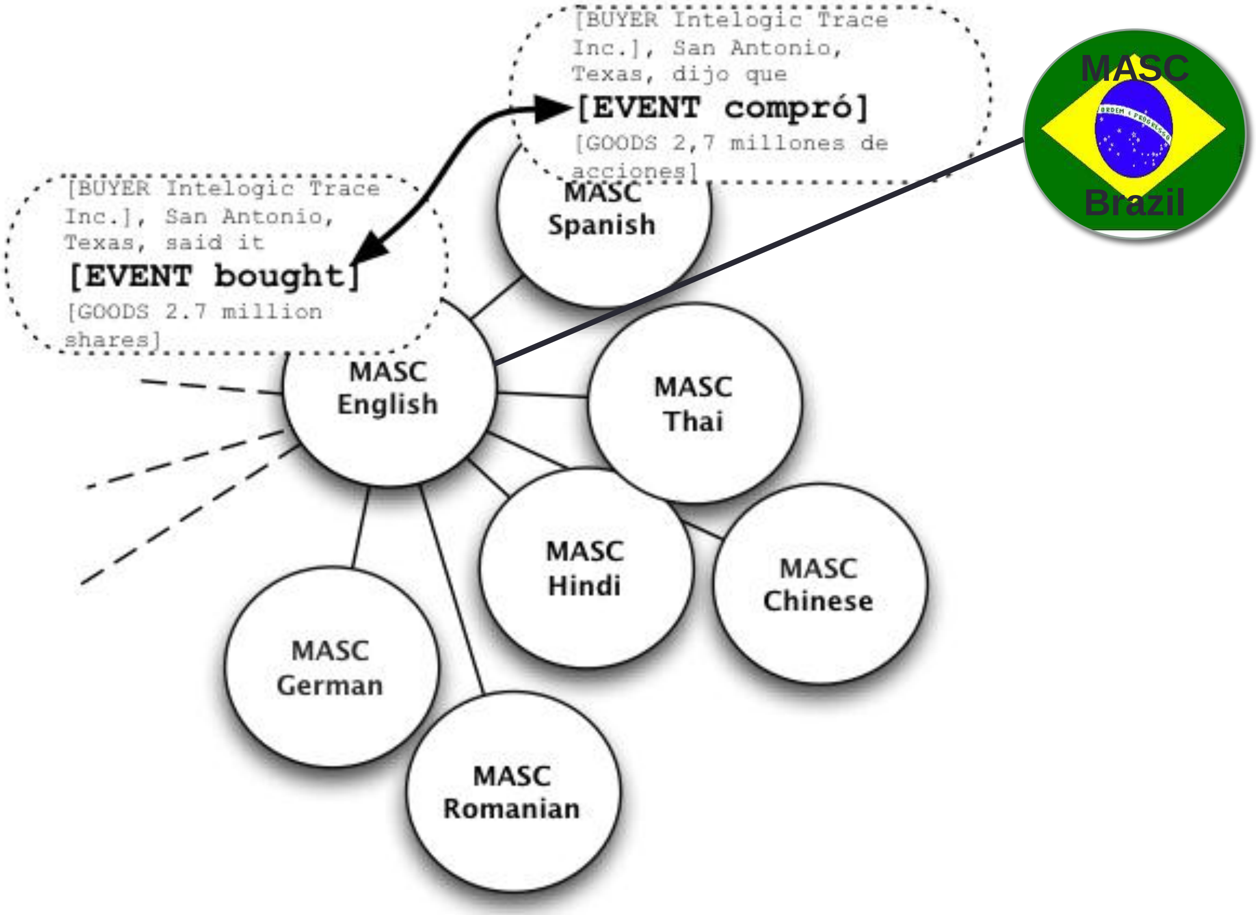
Beyond MASC: Multi-MASC

- Collaborative effort to develop “parallel” corpora to MASC for multiple languages
  - Definition of parallel is flexible: comparable, translated, partial overlap…
  - Comparable annotations
  - Link linguistic phenomena to MASC and/or Multi-MASC corpora in other languages
- Ideal situation for representation in the LLOD cloud

Ide, N. (2012). MultiMASC: An Open Linguistic Infrastructure for Language Research. *Proceedings of the Fifth Workshop on Building and Using Comparable Corpora*, LREC 2012 workshop

The Vision

Why not…


[BUYER Intelogic Trace Inc.], San Antonio, Texas, dijo que
[EVENT compró]
[GOODS 2,7 millones de acciones]
MASC Brazil
[BUYER Intelogic Trace Inc.], San Antonio, Texas, said it
[EVENT bought]
[GOODS 2.7 million shares]
MASC Spanish
MASC English
MASC Thai
MASC Hindi
MASC Chinese
MASC German
MASC Romanian

MultiMASC Challenges

- Finding adequate amounts of fully open data representing the range of genres
  - Difficult!
  - All web documents are copyrighted unless explicitly indicated to be in the public domain or under a specific license such as Creative Commons
    - Cannot redistribute (e.g. with annotations added)
      - Not much good for NLP
      - Great if you do not want to share data and annotations
  - “Share-alike” and GNU Public License (GPL) not good because limit the conditions of redistribution
    - We restrict OANC/MASC texts to public domain or Creative Commons-Attribution (CC-BY)

MultiMASC Challenges

- Finding resources (funding) for development
  - Difficult!
  - For this reason we have outlined an incremental development process
    1. Gather 500K of data
       - If necessary, need not be directly comparable to MASC
    2. Automatically annotate
    3. Validate
    4. Link to other resources
  - At each stage, make openly available

Challenges for Collaborative Resource Development

- Biggest challenge is engaging the community
  - People tend to work in their own environment (easier, more familiar, etc.)
- How to overcome?
  - Organize a shared task that utilizes MASC, OANC, etc.
  - Educate the Computational Linguistics community about advantages of distributed development, LLOD representation, etc.
  - Actively solicit contribution!
  - Others?
  - Suggestions welcome!

Expectation

- Distribution of effort and integration of independent resources such as the OANC, MASC, WordNet, FrameNet, and others will enable progress in resource development
    - Less cost
    - No duplication of effort
    - Greater degree of accuracy and usability
    - Harmonization

Next Steps

- Continually augment MASC and the OANC with contributed annotations from the research community
  - Discourse structure, additional entities, events, opinions/sentiment, etc.
- Continually augment MASC/OANC with new data
  - Requirement for open data only places severe restrictions on what can be included
- Foster and support development of MultiMASC
- Continue sense annotation effort
- PROMOTE COMMITMENT TO OPEN RESOURCES

MASC

- Is currently the largest semantically annotated corpus of English in existence
- Through development of robust annotation procedures, should have a major impact on the speed with which similar resources can be reliably annotated
- WN and FN annotation of the MASC will immediately create a massive multi-lingual resource network
  - Both WN and FN linked to corresponding resources in other languages
  - No existing resource approaches this scope
- Incorporation into LLOD will demonstrate the viability and advantage of representing linguistic resources in the Semantic Web

MASC

- Because it enables merging annotations at different linguistic levels, will
  - facilitate a deeper investigation of interactions among linguistic phenomena
  - contribute to better understanding of the workings of language at the semantic level
- MASC can serve as a model for community effort to develop required methods and resources to further NLP research

Availability

- OANC and MASC available at www.anc.org
- 1st set of MASC data (82K) available now
- Full 500K available within a couple of weeks!
- We encourage (beg for!) contributions of annotations of MASC (manual/validated) and/or OANC (automatic or manual) data for any linguistic phenomenon, in any format
  - We will do the transduction to GrAF
- We offer our tools, expertise, etc. for contributors and developers of MultiMASC corpora

Obtain/Contribute

- http://www.anc.org/MASC
- http://www.anc.org/OANC
- http://www.anc.org:8080/ANC2Go

- Contribute annotations in any format
  - Contact anc@anc.org

THANK YOU