

Open Event Data: Opportunities and Challenges

Philip A. Schrodt

Parus Analytics
schrodt735@gmail.com

Workshop on “Big Data and Death”
University of Wisconsin
7 November 2014

PARUS

ANALYTICS

Event Model: Core Innovation

Once calibrated, real-time event forecasting models can be run entirely without human intervention

- ▶ Web-based news feeds provide a rich multi-source flow of political information in real time
- ▶ Statistical models can be run and tested automatically, and are 100% transparent

In other words, for the first time in human history—quite literally—we have a system that can provide real-time measures of political activity without any human intermediaries

Why Event Data are well suited for predicting political change at short time horizons

- ▶ Structural indicators such as GDP, infant mortality, past or adjacent conflict change too slowly
 - ▶ They nonetheless affect the overall probability
- ▶ Social media indicators change too quickly
 - ▶ Social media appear to give—at best—about a six to twelve hour warning in collective action situations (Carley)
 - ▶ So far, no indications that social media provide reliable indicators of deep social/cultural change: signal-to-noise ratio is very low
 - ▶ Though US government funders are completely obsessed with this at the moment. Tweet that!
- ▶ Newsworthy events are “just right”
 - ▶ And we’ve got the models to prove it
 - ▶ Which is why they are “newsworthy”
 - ▶ Structural indicators either are reflected in the patterns of events, or can be additional covariates

News Story Example: Example: 18 December 2007

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

The Turkish attacks in Dohuk Province on Sunday—involving dozens of warplanes and artillery—were the largest known cross-border attack since 2003. They occurred with at least tacit approval from American officials. The Iraqi government, however, said it had not been consulted or informed about the attacks.

Massoud Barzani, leader of the autonomous Kurdish region in the north, condemned the assaults as a violation of Iraqi sovereignty that had undermined months of diplomacy. “These attacks hinder the political efforts exerted to find a peaceful solution based on mutual respect.”

New York Times, 18 December 2007

http://www.nytimes.com/2007/12/18/world/middleeast/18iraq.html?_r=1&ref=world&oref=slogin
(Accessed 18 December 2007)

TABARI Coding: Lead sentence

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: First event

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: Actors

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: Agent

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: Second event

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: Second event target

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

TABARI Coding: Agent

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD REB

Categorization of Political Interactions

- ▶ Distinct English-language verb phrases:
5,000 to 15,000
(MUC, KEDS, PANDA projects)

- ▶ Micro-level categories
50 to 200
(WEIS, BCOW, IDEA, CAMEO)

- ▶ Macro-level categories
10 to 20
(WEIS, COPDAB, IPB, World Handbook)

WEIS Primary Categories

01	Yield	11	Reject
02	Comment	12	Accuse
03	Consult	13	Protest
04	Approve	14	Deny
05	Promise	15	Demand
06	Grant	16	Warn
07	Reward	17	Threaten
08	Agree	18	Demonstrate
09	Request	19	Reduce Relationship
10	Propose	20	Expel
		21	Seize
		22	Force

CAMEO

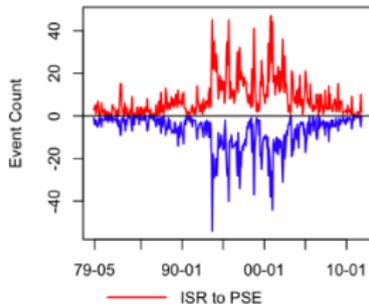
- ▶ 20 primary event categories; around 200 subcategories
- ▶ Based on the WEIS typology but with greater detail on violence and mediation
- ▶ Combines ambiguous WEIS categories such as [WARN/THREATEN] and [GRANT/PROMISE]
- ▶ National actor codes based on ISO-3166 and `CountryInfo.txt`
- ▶ Substate “agents” such as GOV, MIL, REB, BUS
- ▶ Extensive IGO/NGO list

Quad Counts

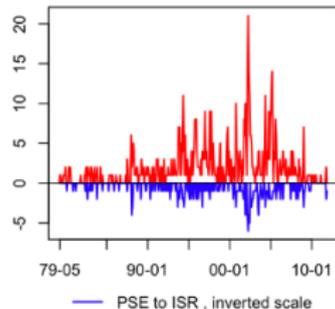
- ▶ Verbal Cooperation (VERCP): The occurrence of dialogue-based meetings (i.e. negotiations, peace talks), statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors. CAMEO categories 01 to 05.
- ▶ Material Cooperation (MATCP): Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc. CAMEO categories 06 to 09.
- ▶ Verbal Conflict (VERCF): A spoken criticism, threat, or accusation, often related to past or future potential acts of material conflict. CAMEO categories 10 to 14.
- ▶ Material Conflict (MATCF): Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc. CAMEO categories 15 to 20.

KEDS Project Levant Data, 1979-2010

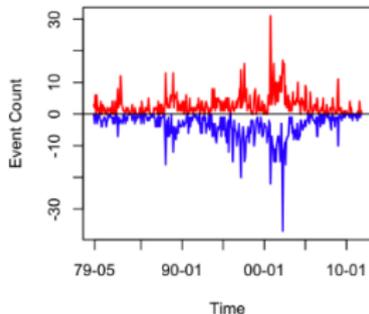
Verbal Cooperation



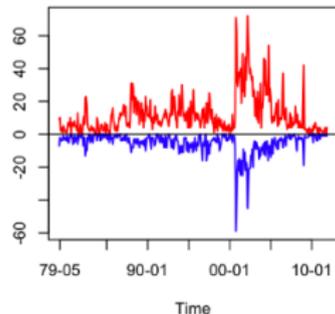
Material Cooperation



Verbal Conflict

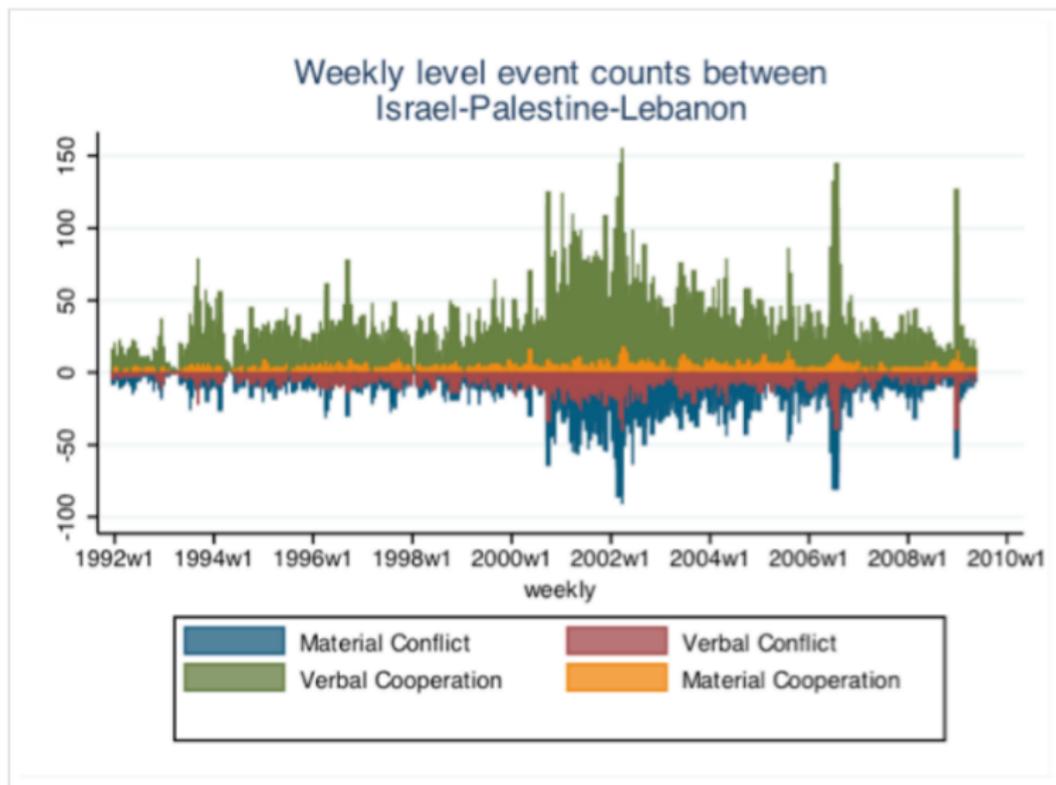


Material Conflict

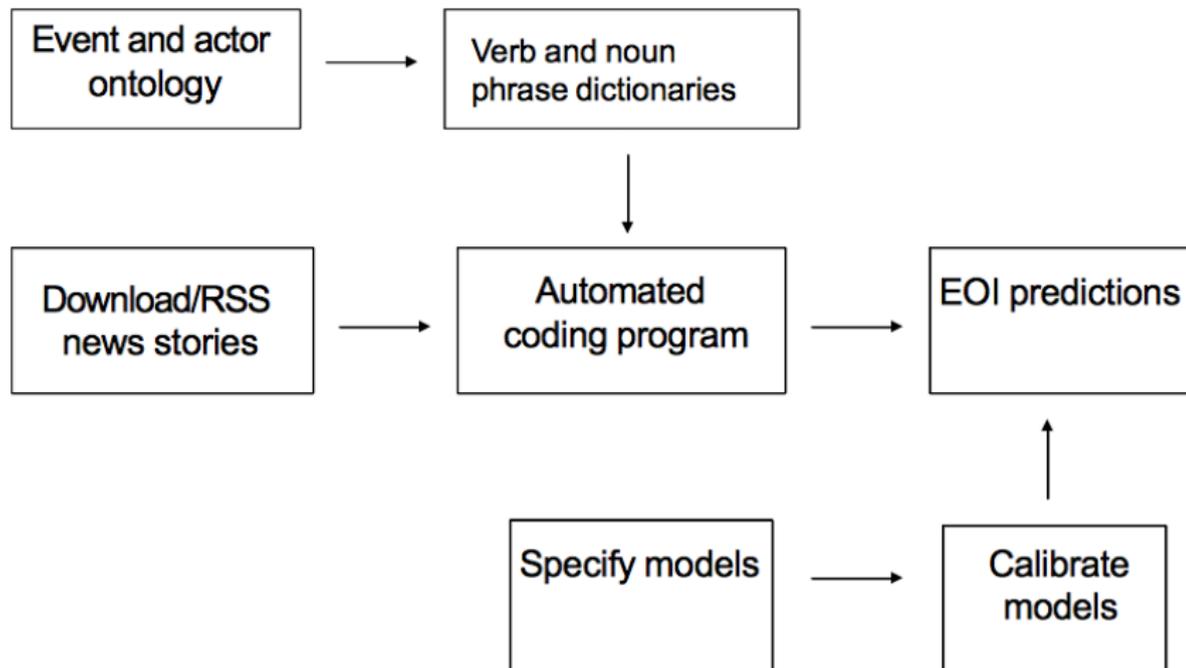


KEDS Project Levant Data, 1992-2010

Visualization by Jay Yonamine



Generating event data



A sort of book on event data

Schrodt and Gerner 2000/2012 *Analyzing International Event Data*, chaps 1-3

<http://eventdata.parusanalytics.com/papers.dir/automated.html>.

We are also hoping to get an textbook-like instructional tool/site and possible some MOOC-like video materials using an open-collaboration environment: this will cover both event data analysis and the toolset.

Challenges to Coding Event Data for Contentious Politics-1

The number of actors who must be identified is substantially greater than the number involved in inter-state events

- ▶ Detailed geographical information—city, region and administrative unit names—may be required
- ▶ Ethnic group names may be important
- ▶ Leadership is less stable—“five minutes of fame”

Coverage in international news sources may be less consistent, with a focus on

- ▶ Major events
- ▶ Periods when a reporter happens to be in the area
- ▶ Events in major cities (or cities with 5-star hotels)

Challenges to Coding Event Data for Contentious Politics-2

Sentences being coded may assume substantial implicit knowledge

- ▶ This is particularly true for full-story coding

In militarized conflicts, large parts of the country may be inaccessible

Activities of unidentified actors may be important: “gunmen killed two journalists. . .”

Modes of reliability in text processing

- **Stability**—the ability of a coder to consistently assign the same code to a given text;
- **Reproducibility**—intercoder reliability;
- **Accuracy**—the ability of a group of coders to conform to a standard.

Source: Weber (1990:17)

In principle, it would be useful to know reproducibility

- ▶ Between coders at different phases of the project
- ▶ Between coders at multiple institutions if the project is decentralized

Advantages of automated coding

- ▶ Fast and inexpensive
- ▶ Transparent: coding rules are explicit in the dictionaries
- ▶ Reproducible: a coding system can be consistently maintained over a period of time without the “coding drift” caused by changing teams of coders.
- ▶ Coding dictionaries can be shared between institutions
- ▶ The coding of individual reports is not affected by the biases of individual coders. Dictionaries, however, can be so affected.
- ▶ It is possible to create rules for difficult technical and cultural vocabulary that is otherwise difficult to learn

Disadvantages of automated coding

- ▶ Automated thematic coding has problems with disambiguation
- ▶ Automated syntactic coding using shallow parsing makes errors on complex sentences by incorrectly identifying the object of the sentence.
- ▶ Requires a properly formatted, machine-readable source of text, therefore older paper and microfilm sources are difficult to code.
- ▶ Development of new coding dictionaries is time-consuming—KEDS/PANDA initial dictionary development required 2-labor-years. (Modification of existing dictionaries, however, requires far less effort)

Human vs Machine Coding: Summary

Advantage to human coding

- ▶ Small data sets
- ▶ Data coded only one time at a single site
- ▶ Existing dictionaries cannot be modified
- ▶ Complex sentence structure
- ▶ Metaphorical, idiomatic, or time- dependent text
- ▶ Money available to fund coders and supervisors

Advantage to machine coding

- ▶ Large data sets
- ▶ Data coded over a period of time or across projects
- ▶ Existing dictionaries can be modified
- ▶ Simple sentence structures
- ▶ Literal, present-tense text
- ▶ Money is limited

But fundamentally, comparisons with human coding are irrelevant if one is coding over a billion sentences and updating at the rate of 100,000 stories per day.

EL:DIABLO

Event Location: Dataset in a Box, Linux Option

- ▶ Full modular open-source pipeline to produce daily event data from web sources
- ▶ Scraper from white-list of RSS feeds and web pages
- ▶ Event coding from PETRARCH but other coders easily added to the pipeline
- ▶ Conventional deduplication keeping URLs of all duplicates
- ▶ Additional feature detectors are easily added
- ▶ Designed for implementation in Linux cloud (e.g. Linode: \$20/month)

Event data in a box, basically.

📄 20 commits

🌿 2 branches

📦 0 releases

👤 1 contributor

branch: **master** ▾[eldiablo](#) / [+](#)

Tinkering.

[Johnb30](#) authored 16 days agolatest commit [933deaafe0](#) [.gitignore](#)

It all works now

2 months ago

[LICENSE](#)

Initial commit

2 months ago

[README.md](#)

Tinkering.

16 days ago

[Vagrantfile](#)

Adding files.

2 months ago

[bootstrap.sh](#)

It all works now

2 months ago

[crontab.txt](#)

Fix typo in crontab.txt

2 months ago

[README.md](#)

<> Code

🔍 Issues 0

📄 Pull Requests 0

📊 Pulse

📊 Graphs

🌐 Network

HTTPS clone URL

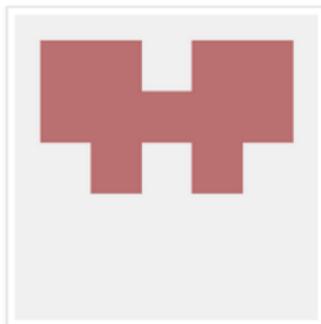
<https://github.com> You can clone with [HTTPS](#) or [Subversion](#). 📄

Clone in Desktop

Download ZIP

PETRARCH

- ▶ Written in Python, in contrast to the C++ TABARI
- ▶ Full parsing using the Penn Treebank format and Stanford Core NLP. This handles the noun/verb/adjective disambiguation that accounts for much of the size of the TABARI dictionaries
- ▶ Synonym sets from WordNet
- ▶ Identifies actors even if they are not in the dictionaries
- ▶ Extendible through program “hooks”: “issues” facility
- ▶ Codes at about 150 sentences per second, about a tenth the speed of TABARI but cluster computing is now readily available
- ▶ Problem: TABARI dictionaries—based on shallow parsing—do not always translate well to the higher precision of full parsing



Philip Schrodt

eventdata

📍 University Park, PA 16801
USA

🌐 <http://eventdata.psu.edu>

🕒 Joined on Feb 22, 2012

3

public repos

5

members

[📁 Repositories](#)[👤 Members](#)[All Sources](#)

PETRARCH

Python-language successor to the TABARI event data program

Last updated 10 days ago



PETRARCH_ec2

PETRARCH version optimized to run on an Amazon EC2 cluster

Last updated 4 months ago



Computational-Approaches

Last updated 2 years ago

Improving JABARI Accuracy

- ▶ TABARI baseline: 56% precision, 54% recall
- ▶ Add Open-NLP Penn TreeBank parser: 68% precision, 35.4% recall
- ▶ Add GATE-Annie noun phrase synonyms, pronoun coreferencing, and default location agent resolution: 77% precision, 66.5% recall

Why Python?

- ▶ Open source (of course...tools want to be free...)
- ▶ Standardized across platforms and widely available/documented
- ▶ Automatic memory management (unlike C/C++)
- ▶ Generally more coherent than perl, particularly when dealing with large programs
- ▶ Text oriented rather than GUI oriented (unlike Java)
- ▶ Extensive libraries but these are optional (unlike Java): seems to be generating very substantial network effects
- ▶ C/C++ code can be easily integrated in high-performance applications
- ▶ Tcl can be used for GUI

Sources for historical texts

- ▶ LDC Gigaword 2000-2010; easily licensed
- ▶ Cline Center, University of Illinois at Urbana-Champaign (being coded as we speak)
- ▶ Usual proprietary sources: we are making some real progress negotiating affordable licenses now
- ▶ Collective resources: in the US, coded data on facts does not inherit the IP constraints of the source
- ▶ Discussion of legal issues for US:
<http://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/>

Open Event Data Alliance

- ▶ Institutionalize event data following the model of CRAN and many other decentralized open collaborative research groups: these turn out to be common in most research communities
- ▶ Provide at least one source of daily updates with 24/7/365 data reliability. Ideally, multiple such data sets rather than “one data set to rule them all”
- ▶ Establish common standards, formats, and best practices
- ▶ Open source, open collaboration, open access

Extending the event ontologies

CAMEO and IDEA were derived from earlier Cold War event ontologies (WEIS, COPDAB, World Handbook) and consequently miss substantial amounts of political behavior that is currently relevant.

- ▶ natural disaster
- ▶ disease
- ▶ criminal activity
- ▶ financial activity
- ▶ refugees and related humanitarian issues
- ▶ human rights violations
- ▶ electoral and parliamentary activity

Reference: Philip A. Schrodtt and Benjamin Bagozzi. 2012. “Detecting the Dimensions of News Reports using Latent Dirichlet Allocation Models.” European Political Science Association meetings, Berlin. <http://eventdata.parusanalytics.com/papers.html>.

We may finally get the ICEWS data!

People involved say this may be available in January 2015

- ▶ BBN/Lockheed [proprietary] Serif coder
- ▶ CAMEO ontology
- ▶ Factiva based; density is 2000 to 4000 events per day
- ▶ Coverage probably 2000-present with a rolling one-year embargo, released monthly
- ▶ Data includes a version of the sentence generating the event (!)
- ▶ Actor dictionaries will be released
- ▶ Coverage probably 2000-present with a rolling one-year embargo, updated monthly
- ▶ Geolocated, though not clear how useful this is

Advantages of the CoreNLP parsing compared to TABARI shallow parsing

- ▶ Reduces incorrect identification of direct objects, which messes up source identification
- ▶ Provides noun/verb/adjective disambiguation: many words in English can be used in all three modes:
 - ▶ “A protest occurred on Sunday” [noun]
 - ▶ “Demonstrators protested” [verb]
 - ▶ “Marchers carried protest signs” [adjective]
- ▶ Identification of all named entities through noun phrases:
 - ▶ TABARI required actor to be in dictionaries.
 - ▶ PETRARCH will always pull these out whenever they occur in the source or target position;
 - ▶ The result unidentified cases can be separately processed with named-entity-resolution (NER) software
- ▶ More sophisticated co-referencing of pronouns and other references, particularly across sentences

Stanford CoreNLP parse tree

```
<EventID date="19950103" id="DEMO-04" category="DEMO">
<!-- [Paired events: LEFT_ generates a "visit" and "receive visit" events] -->
<EventCoding sourcecode="DAG" targetcode="GON" eventcode="032">
<EventCoding sourcecode="GON" targetcode="DAG" eventcode="033">
Dagolath's first Deputy Prime Minister Telemar left for
Minas Tirith on Wednesday for meetings of the joint transport
committee with Arnor, the Dagolathi news agency reported.
(ROOT
 (S
  (S
   (NP
    (NP (NNP Dagolath) (POS 's))
     (ADJP (JJ first))
     (NNP Deputy) (NNP Prime) (NNP Minister) (NNP Telemar))
   (VP (VBD left)
    (PP (IN for)
     (NP
      (NP (NNP Minas) (NNP Tirith))
       (PP (IN on)
        (NP (NNP Wednesday))))))
    (PP (IN for)
     (NP
      (NP (NNS meetings))
       (PP (IN of)
        (NP
         (NP (DT the) (JJ joint) (NN transport) (NN committee))
          (PP (IN with)
           (NP (NNP Arnor))))))))))
    (, ,)
   (NP (DT the) (NNP Dagolathi) (NN news) (NN agency))
   (VP (VBD reported))
  (. .)))
```


Problems PETRARCH/CoreNLP does not solve

- ▶ Word-sense disambiguation
 - ▶ "attack": physical or verbal?

WordNet word senses: “attack”

Noun

- S: (n) **attack**, onslaught, **onset**, **onrush** ((military) an offensive against an enemy (using weapons)) *"The attack began at dawn"*
- S: (n) **attack** (an offensive move in a sport or game) *"They won the game with a 10-hit attack in the 9th Inning"*
- S: (n) fire, **attack**, **flak**, **flack**, **blast** (intense adverse criticism) *"Clinton directed his fire at the Republican Party"; "the government has come under attack"; "don't give me any flak"*
- S: (n) approach, **attack**, **plan of attack** (ideas or actions intended to deal with a problem or situation) *"his approach to every problem is to draw up a list of pros and cons"; "an attack on inflation"; "his plan of attack was misguided"*
- S: (n) **attack**, **attempt** (the act of attacking) *"attacks on women increased last year"; "they made an attempt on his life"*
- S: (n) **attack**, **tone-beginning** (a decisive manner of beginning a musical tone or phrase)
- S: (n) **attack** (a sudden occurrence of an uncontrollable condition) *"an attack of diarrhea"*
- S: (n) **attack** (the onset of a corrosive or destructive process (as by a chemical agent)) *"the film was sensitive to attack by acids"; "open to attack by the elements"*
- S: (n) **attack** (strong criticism) *"he published an unexpected attack on my work"*

Verb

- S: (v) **attack**, **assail** (launch an attack or assault on; begin hostilities or start warfare with) *"Hitler attacked Poland on September 1, 1939 and started World War II"; "Serbian forces assailed Bosnian towns all week"*
- S: (v) **attack**, **round**, **assail**, **lash out**, **snipe**, **assault** (attack in speech or writing) *"The editors of the left-leaning paper attacked the new House Speaker"*
- S: (v) **attack**, **aggress** (take the initiative and go on the offensive) *"The Serbs attacked the village at night"; "The visiting team started to attack"*
- S: (v) **assail**, **assault**, **set on**, **attack** (attack someone physically or emotionally) *"The mugger assaulted the woman"; "Nightmares assaulted him regularly"*
- S: (v) **attack** (set to work upon; turn one's energies vigorously to a task) *"I attacked the problem as soon as I got out of bed"*
- S: (v) **attack** (begin to injure) *"The cancer cells are attacking his liver"; "Rust is attacking the metal"*

Problems PETRARCH/CoreNLP does not solve

- ▶ Word-sense disambiguation
 - ▶ "attack": physical or verbal?
 - ▶ "head" has about 65 different meanings in English, ranging from a leadership designation to a marine toilet.

WordNet word senses: “head”

Noun

- S: (n) **head**, **caput** (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head** (a single domestic animal) *"200 head of cattle"*
- S: (n) **mind**, **head**, **brain**, **psyche**, **nous** (that which is responsible for one's thoughts and feelings; the seat of the faculty of reason) *"his mind wandered"; "I couldn't get his words out of my head"*
- S: (n) **head**, **chief**, **top dog** (a person who is in charge) *"the head of the whole operation"*
- S: (n) **head** (the front of a military formation or procession) *"the head of the column advanced boldly"; "they were at the head of the attack"*
- S: (n) **head** (the pressure exerted by a fluid) *"a head of steam"*
- S: (n) **head** (the top of something) *"the head of the stairs"; "the head of the page"; "the head of the list"*
- S: (n) fountainhead, **headspring**, **head** (the source of water from which a stream arises) *"they tracked him back toward the head of the stream"*
- S: (n) **head**, **head word** ((grammar) the word in a grammatical constituent that plays the same grammatical role as the whole constituent)
- S: (n) **head** (the tip of an abscess (where the pus accumulates))
- S: (n) **head** (the length or height based on the size of a human or animal head) *"he is two heads taller than his little sister"; "his horse won by a head"*
- S: (n) **capitulum**, **head** (a dense cluster of flowers or foliage) *"a head of cauliflower"; "a head of lettuce"*
- S: (n) principal, school principal, **head teacher**, **head** (the educator who has executive authority for a school) *"she sent unruly pupils to see the principal"*
- S: (n) **head** (an individual person) *"tickets are \$5 per head"*
- S: (n) **head** (a user of (usually soft) drugs) *"the office was full of secret heads"*
- S: (n) **promontory**, **headland**, **head**, **foreland** (a natural elevation (especially a rocky one that juts out into the sea))
- S: (n) **head** (a rounded compact mass) *"the head of a comet"*
- S: (n) **head** (the foam or froth that accumulates at the top when you pour an effervescent liquid into a container) *"the beer had a large head of foam"*
- S: (n) **forefront**, **head** (the part in the front or nearest the viewer) *"he was in the forefront"; "he was at the head of the column"*
- S: (n) pass, **head**, **strait** (a difficult juncture) *"a pretty pass"; "matters came to a head yesterday"*
- S: (n) **headway**, **head** (forward movement) *"the ship made little headway against the gale"*
- S: (n) **point**, **head** (a V-shaped mark at one end of an arrow pointer) *"the point of the arrow was due north"*
- S: (n) **question**, **head** (the subject matter at issue) *"the question of disease merits serious discussion"; "under the head of minor Roman poets"*

WordNet word senses: “head” continued

Noun

- S: (n) heading, **header**, **head** (a line of text serving to indicate what the passage below it is about) *“the heading had little to do with the text”*
- S: (n) **head** (the rounded end of a bone that fits into a rounded cavity in another bone to form a joint) *“the head of the humerus”*
- S: (n) **head**, **caput** (the upper part of the human body or the front part of the body in animals; contains the face and brains) *“he stuck his head out the window”*
- S: (n) **head** (that part of a skeletal muscle that is away from the bone that it moves)
- S: (n) **read/write head**, **head** ((computer science) a tiny electromagnetic coil and metal pole used to write and read magnetic patterns on a disk)
- S: (n) **head** ((usually plural) the obverse side of a coin that usually bears the representation of a person's head) *“call heads or tails!”*
- S: (n) **head** (the striking part of a tool) *“the head of the hammer”*
- S: (n) **head** ((nautical) a toilet on board a boat or ship)
- S: (n) **head** (a projection out from one end) *“the head of the nail”, “a pinhead is the head of a pin”*
- S: (n) **drumhead**, **head** (a membrane that is stretched taut over a drum)

Verb

- S: (v) **head** (to go or travel towards) *“where is she heading?”; “We were headed for the mountains”*
- S: (v) **head**, **lead** (be in charge of) *“Who is heading this project?”*
- S: (v) **lead**, **head** (travel in front of; go in advance of others) *“The procession was headed by John”*
- S: (v) **head**, **head up** (be the first or leading member of (a group) and excel) *“This student heads the class”*
- S: (v) **steer**, **maneuver**, manoeuvre, manoeuvre, **direct**, **point**, **head**, **guide**, **channelize**, **channelise** (direct the course; determine the direction of travelling)
- S: (v) **head** (take its rise) *“These rivers head from a mountain range in the Himalayas”*
- S: (v) **head** (be in the front of or on top of) *“The list was headed by the name of the president”*
- S: (v) **head** (form a head or come or grow to a head) *“The wheat headed early this year”*
- S: (v) **head** (remove the head of) *“head the fish”*

Problems PETRARCH/CoreNLP does not solve

Detailed development (and extension) of the CAMEO categories and dictionaries

- ▶ CAMEO was developed to study mediation, not as a general-purpose coding ontology
- ▶ Converting the TABARI dictionaries from WEIS to CAMEO took about three academic-research-project-years
- ▶ This is mundane, sloggy, labor intensive task on the same scale as a large human-coded data project
- ▶ it is not the sort of big data sexy topic that funders are ready to throw gobs of open-source/open-access money at.

WordNet-based dictionaries

PRINCETON UNIVERSITY

WordNet

A lexical database for English



What is WordNet?

- What is WordNet?
- People
- News
- Use WordNet online
- Download
- Citing WordNet
- License and commercial use
- Related projects
- WordNet documentation
- Publications
- Frequently Asked Questions

Current News

[George A. Miller](#), who began the WordNet project in the mid-1980s, passed away on July 22, 2012 at the age of 92. You can read his obituary [here](#).

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the [browser](#). WordNet is also freely and publicly available for [download](#). WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

We appreciate your comments and suggestions, especially when they are constructive and help us improve WordNet. We get numerous questions regarding topics that are addressed on our [FAQ](#) page. Before you [email the WordNet team](#), please [check the FAQ](#) first to see if the answer is already there. If you have a problem or question regarding something you downloaded from the ["distilled projects"](#) page, you must contact the developer directly.

Our staff examines all mail and tries to make appropriate database changes, but we hope you understand that due to time and staff constraints we cannot always respond.

Please note that changes made to the database are not reflected until a new version of WordNet is

WordNet-based dictionaries

- ▶ Verb dictionaries have been completely reorganized around *WordNet* synonym sets (“synsets”)
- ▶ Verb-phrase patterns include synsets for common objects such as currency, weapons and quantities
- ▶ “Agents” dictionary for common nouns—for example “police”, “soldiers”, “president”—includes all *WordNet* synsets
- ▶ Dictionaries will be reformatted into a JSON data structure
- ▶ Additional dictionary enhancements carried forward from TABARI 0.8
 - ▶ regular noun and verb endings
 - ▶ all irregular verb forms
 - ▶ improved dictionaries for militarized non-state actors

Portugal vs. Israel???

Portugal to Deploy Untried Defence Against Israel

By REUTERS

Published: October 9, 2013 at 12:50 PM ET

Portugal vs. Israel???

Portugal to Deploy Untried Defence Against Israel

By REUTERS

Published: October 9, 2013 at 12:50 PM ET

LISBON — Portugal are close to securing at least a World Cup playoff place but their untested back four will be under scrutiny against Israel in Friday's Group F qualifier at the Alvalade stadium (1945 GMT).



League Scoreboards

- Major League Soccer
- English Premier League
- Champions League
- Bundesliga
- Serie A | La Liga

Coach Paulo Bento must patch up his defence after a string of injuries affected right backs Joao Pereira and Miguel Lopes, as well as centre back Bruno Alves.

Left back Fabio Coentrao is also out suspended as Portugal, a point behind leaders Russia, look set to miss out on

automatic qualification unless the Russians slip up against Luxembourg and Azerbaijan in their remaining fixtures.

Third-placed Israel are five points behind Portugal.

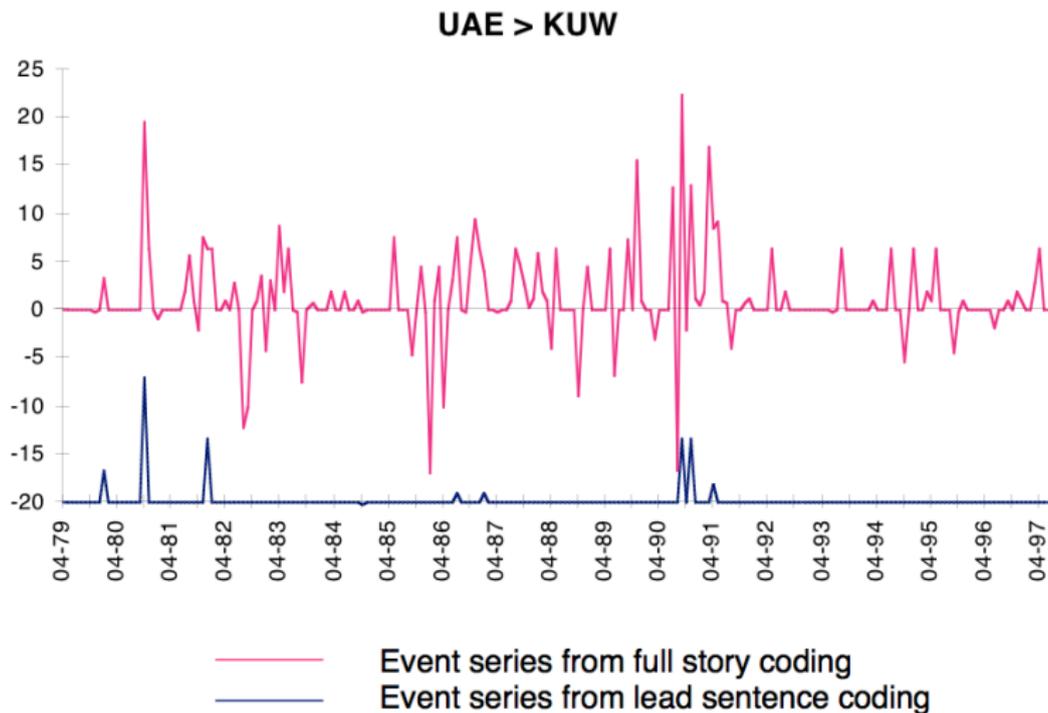
"We simply must win. We know Israel's strong points, how much we suffered there and how good their finishers and counter attacks are," striker Hugo Almeida told reporters in the medieval fortress town of Obidos, where Portugal are training.

	FACEBOOK
	TWITTER
	GOOGLE+
	SAVE
	E-MAIL
	SHARE
	PRINT
	REPRINTS

“fortress town”: thanks...

Source:<http://www.nytimes.com/reuters/2013/10/09/sports/soccer/09reuters-soccer-portugal.html>

Full story vs. lead sentence coding [KEDS]



Named Entity Recognition/Resolution

- ▶ Locating and classifying phrases into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- ▶ Examples:
<http://www ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html>.
- ▶ No general solution; approaches tend to be either
 - ▶ Rule and dictionary based, which requires manual development
 - ▶ Sequence-based machine-learning methods, specifically conditional random fields. These require an extensive set of marked-up examples
- ▶ *Name resolution* involves either
 - ▶ Differentiating two distinct entities which have the same name: “President Bush”
 - ▶ Combining multiple names of the same entity”
“Obamacare” and “Affordable Care Act”
- ▶ Network models which associate a particular use of the

Actors are distributed approximately rank-size

Phrase	Proportion of events
China	11.86%
United States	11.56%
Russian Federation	8.43%
Japan	7.99%
North Korea	5.33%
India	5.24%
South Korea	3.45%
Chinese	3.44%
UN	3.14%
Taiwan	3.13%
Pakistan	3.10%
Thailand	2.88%
Australia	2.48%
Iraq	2.23%
United Kingdom	2.08%
Indonesia	1.96%

Observation: the “short snout” is much more important than the “long tail.”

Actor distribution: sample of the tail

Phrase	Proportion of events
Hamid Karzai	0.01%
President (Angola)	0.01%
Yang Hyong Sop	0.01%
Kashmir State	0.01%
Ehud Olmert	0.01%
Police (Sri Lanka)	0.01%
Vojislav Kostunica	0.01%
Commerce Minist (India)	0.01%
Parliament (Iran)	0.01%
President (Yemen)	0.01%
Foreign Minist (Netherlands)	0.01%
Director General (IAEA)	0.01%
Liu Qi	0.01%
Yang Jiechi	0.01%
Business (Hong Kong)	0.01%
President (Namibia)	0.01%
Police (China)	0.01%
Business (France)	0.01%

Goldstein Scale [WEIS]

010:	[1.0]	YIELD	110:	[-4.0]	REJECT
011:	[0.6]	SURRENDER	111:	[-4.0]	TURN DOWN
012:	[0.6]	RETREAT	112:	[-4.0]	REFUSE
013:	[2.0]	RETRACT	113:	[-5.0]	DEFY LAW
014:	[3.0]	ACCOMODATE, CEASEFIRE	170:	[-6.0]	THREATEN
015:	[5.0]	CEDE POWER	171:	[-4.4]	UNSPECIFIED THREAT
020:	[0.0]	COMMENT	172:	[-5.8]	NONMILITARY TRHEAT
021:	[-0.1]	DECLINE COMMENT	173:	[-7.0]	SPECIFIC THREAT
022:	[-0.4]	PESSIMISTIC COMMENT	174:	[-6.9]	ULTIMATUM
023:	[-0.2]	NEUTRAL COMMENT	220:	[-9.0]	FORCE
024:	[0.4]	OPTIMISTIC COMMENT	221:	[-8.3]	NONINJURY DESTRUCTION
070:	[7.0]	REWARD	222:	[-8.7]	NONMIL DESTRUCTION
071:	[7.4]	EXTEND ECON AID	223:	[-10.0]	MILITARY ENGAGEMENT
072:	[8.3]	EXTEND MIL AID			
073:	[6.5]	GIVE OTHER ASSISTANCE			

Problems with the Goldstein scale

- ▶ It started out quite arbitrary, and the CAMEO versions are even worse
- ▶ It tends to be dominated by violence events, which mask low levels of cooperative events
- ▶ It correlates highly with the event count, and in fact simple event counts do almost as well, similar to the result that unweighted equations do well
- ▶ The data are nominal!: get over it

Additional work to be done

Specialized data sets

- ▶ Protest
 - ▶ Size
 - ▶ Topic[s]
 - ▶ Sponsor[s]
 - ▶ Response of authorities[s]
 - ▶ Location resolved below the city level
- ▶ Monitoring/situational awareness
 - ▶ Minimize the false positive rate
 - ▶ Quad-category only
 - ▶ Specialized categories only, e.g. events possibly related to climate change

Major issue: how can we integrate dictionaries produced at multiple sites to maximize the total coverage?

Increasing the speed and efficiency of dictionary development

- ▶ NER systems for near-real-time updating of actors and open collaboration on maintenance of major actor dictionaries
- ▶ Automated identification of new verb phrases: we've never tried this
- ▶ Cloud-sourcing elements of dictionary development and validation
 - ▶ CAMEO is almost certainly too complex for Mechanical Turk, but might be sourced to more professional sites such as Elance and ODesk.
 - ▶ This is more costly than MT but still would scale and is probably cheaper and preferable to a traditional undergraduate coding farm
- ▶ Establishing a “ground truth” validation set covering all of the CAMEO categories
- ▶ Standardization of religion, ethnic groups and militarized non state actors

Expanding local coverage

- ▶ Locating sources which are either open access or have non-predatory licensing arrangements
 - ▶ Event-to-source “drill-down” is a very high priority
 - ▶ Sources need to be shared across projects even if they are not open
 - ▶ al-Jazeera?
 - ▶ “Wikinews”?
- ▶ Non-English sources, probably through Google Translate or a comparable system
- ▶ Location-specific dictionaries for actors and events
- ▶ Utilize NGO sources to the extent that this is ethical and secure

Questions?

Email: schrodt735@gmail.com

Slides:

<http://eventdata.parusanalytics.com/presentations.html>

Software: <https://openeventdata.github.io/>

Papers:

<http://eventdata.parusanalytics.com/papers.html>