Phoenix Event Data Set Documentation *

Philip A. Schrodt
Parus Analytical Systems
schrodt735@gmail.com

Version 0.5 [DRAFT] : February 6, 2014

Text Sources

Phoenix uses a white-list of sources that are accessed via RSS feed. Table 1 shows the current list of sources and their identifiers.

Text Filtering

The following filtering rules applied to the texts

1. Only the first 4 sentences of the story are coded. Full-story coding tends to generate a large number of false positives due to references to historical events and counter-factual speculation, whereas the early sentences generally provide succinct literal statements about the event. Many analyses use only the "lede" (first) sentence: a data set containing only lede coding can be obtained by filtering for the string "`-1`" in the source identifier (field 10).

2. Sentences with fewer than 120 characters were skipped: these usually do not contain reports of political interactions and instead are editorial artifacts (e.g. photo captions), tabular material (e.g. sports scores or prices), or fragments that have resulted from incorrect sentence delineation.

*Funding for development of the Phoenix system was provided in part by the National Science Foundation Political Science and Methodology, Measurement and Statistics programs. (SES-1004414, SES-1259190)

Table 1: Sources

IRI	http://www.irinnews.org/irin.xml
ALA	http://allafrica.com/tools/headlines/rdf/latest/headlines.rdf
RFI	http://www.english.rfi.fr/last_24h/rss
TZA	http://www.todayszaman.com/104.rss
AAK	http://feeds.feedburner.com/AlAkhbarEnglish?format=xml
ICR	http://www.insightcrime.org/news/feed
AJA	http://america.aljazeera.com/content/ajam/articles.rss
CSM	http://rss.csmonitor.com/feeds/world?format=xml
CSM	http://rss.csmonitor.com/feeds/usa?format=xml
CSM	http://rss.csmonitor.com/feeds/politics?format=xml
GOO	https://news.google.com/?output=rss [1]
NYT	http://rss.nytimes.com/services/xml/rss/nyt/World.xml
REU	http://feeds.reuters.com/Reuters/worldNews
BBC	http://feeds.bbci.co.uk/news/world/rss.xml
UPI	http://rss.upi.com/news/emerging_threats.rss
XIN	http://www.xinhuanet.com/english/rss/worldrss.xml
VOA	http://www.voanews.com/api/epiqq
VOA	http://www.voanews.com/api/z-$otevtiq
VOA	http://www.voanews.com/api/zo$o_egviy
VOA	http://www.voanews.com/api/zr$opeuvim
VOA	http://www.voanews.com/api/zj$oveytit
VOA	http://www.voanews.com/api/zoripegtim
VOA	http://www.voanews.com/api/zji-veyj-v
FRA	http://www.france24.com/en/americas/rss/
FRA	http://www.france24.com/en/middle-east/rss
FRA	http://www.france24.com/en/asia-pacific/rss/
FRA	http://www.france24.com/en/africa/rss
GUA	http://www.theguardian.com/world/europe/roundup/rss
GUA	http://www.theguardian.com/world/americas/roundup/rss
GUA	http://feeds.theguardian.com/theguardian/world/china/rss
GUA	http://www.theguardian.com/world/africa/roundup/rss
GUA	http://www.theguardian.com/world/southandcentralasia/roundup/rss
YAH	http://in.news.yahoo.com/rss/asia

1. If one of the regular sources is identified in a Google feed, that source abbreviation is used rather than GOO.

3. Sentences beginning with quotation marks are skipped: direct quotations frequently cannot be correctly coded because of implicit (rather than literal) content and are best skipped.

Coding Engine

TABARI 0.8.4. `http://eventdata.parusanalytics.com/software.dir/tabari.html`

Coding dictionaries are listed in the internal documentation of the daily data sets.

Coding Ontology

CAMEO for events and actors: `http://eventdata.parusanalytics.com/data.dir/cameo.html`

Data Format

The event data is in a tab-delimited text file with Unix line endings. The fields have the following information

1. Date in YYMMDD format
2. 3-character country code for source actor: this is usually an ISO-3166 Alpha-3 code but can also include CAMEO non-state actor codes such as IGO, NGO, MNC, and IMG. See Note 1
3. 3-character CAMEO agent code for the source [may be empty]
4. CAMEO secondary agent code[s] for the source. This is usually 3 characters but may consistent of multiple 3-character code segments, this field also may be empty
5. 3-character country code for the target
6. 3-character CAMEO agent code for the target [may be empty]
7. CAMEO secondary agent code[s] for the target [may be empty]
8. CAMEO event code
9. Short description of the event code

10. Event source identifier: see Note 2
11. URL for source text
12. Number of duplicates
13. Duplicate source list [may be empty]: see Note 3

Notes:

1. The initial lines of the daily files contain documentation about the coding session, including the time and dictionaries used. These are identified by the sequence `DOC DOC 999`. An example:

```
140204 DOC DOC 999 Data generated by TABARI version 0.8.4b1
140204 DOC DOC 999 Thu 06 Feb 2014 16:43 EST
140204 DOC DOC 999 Session 8; Coder PHOX
140204 DOC DOC 999 Phoenix daily update file
140204 DOC DOC 999 Produced by the Open Event Data Alliance: http://openeventdata.org
140204 DOC DOC 999 License: MIT License, http://opensource.org/licenses/MIT
140204 DOC DOC 999 <verbsfile>  CAMEO.091003.master.verbs
140204 DOC DOC 999 <actorsfile>  nouns_adj_null.110124.txt
140204 DOC DOC 999 <actorsfile>  Phoenix.Countries.140130.actors.txt
140204 DOC DOC 999 <actorsfile>  Phoenix.Internatnl.140130.actors.txt
140204 DOC DOC 999 <actorsfile>  Phoenix.MNSA.140131.actors.txt
140204 DOC DOC 999 <agentfile>    Phoenix.140127.agents.txt
140204 DOC DOC 999 <optionsfile> pipeline.options
```

2. The source identifier is of the form `AAA-NNNN-S` where `AAA` is the source identifier from Table 1, `NNNN` is a zero-filled sequence number of the story in the download, and `S` is the ordinal number of the sentence within the story. The ordinal number is the order in the original report, not the order adjusted for the number of sentences remaining after filtering. The sequence number is generally not meaningful since the ordering of the text is simply determined by the times they are found by the system.

2. The duplicate source list is a series of blank-delimited tuples consisting of a source abbreviation and the number of times a story generating an identical event was found in that source. So for example if two Reuters texts and one BBC text also generated this event,

this field would read `REU 2 BBC 1`. See discussion below on the de-duplication process.

Duplicates

Phoenix uses a “one-a-day filter” to eliminate duplicate reports: only a single record with the same combination of source-target-event is allowed per day. This eliminates multiple reports of the same event—for example a car bombing—by the same source or by multiple sources. The number of duplicates is reported in field 12 and tends to correlate with the importance of the event as assessed by the sources.

In order to reduce the size of the event data file, Phoenix provides only a single URL for the first instance of a duplicated event, then provides the URLs for the duplicates in a separate file: these are keyed to the event in the data set. The primary URL simply corresponds to the first story encountered with a particular source-target-event sequence: it does not indicate the first or most credible report.

Example

[URLs have been abbreviated to fit on the page; full URLs are provided in the actual index]:

```
140204 RUS SYRGOV 012 Make pessimistic comment REU-0021-2 http://news.goog
REU 1 REU-0022-2 http://feeds.reu
TZA 1 TZA-0004-3 http://www.today

140204 AFGREB AFGREB 036 Express intent to negotiate GUA-0016-2 http://www.thegu
GUA 1 GUA-0016-2 http://www.thegu

140204 AUS NZLGOV 040 Consult UPI-0002-1 http://www.upi.c
UPI 1 UPI-0013-1 http://www.upi.c
```