Current State of Automated Event Data Coding

Philip A. Schrodt
Penn State/PRIO
schrodt@psu.edu
http://eventdata.psu.edu

Presented at the Workshop on
Transforming Political Violence
Peace Research Institute Oslo
8-9 March 2012

Changes 2005-2010

- Major expansion of text available on the Web
- Widespread development of open source natural language (NLP) processing software
- DARPA Integrated Conflict Early Warning System (ICEWS): $38M

Existing global data sets

- VRA: 1991-2010, IDEA ontology, probably available for $50K
- Lockheed ICEWS: 1998-2011, CAMEO ontology, supposedly available for research use “soon”
- Nardulli/Leetaru: 1945[!]-2010, SPEED ontology, also “soon”

Existing global data sets

- VRA: 1991-2010, IDEA ontology, probably available for $50K
- Lockheed ICEWS: 1998-2011, CAMEO ontology, supposedly available for research use “soon”
- Nardulli/Leetaru: 1945[!]-2010, SPEED ontology, also “soon”
- Leetaru: 1850[!?!]-2010, CAMEO state and agent level coding

Coding engines

- Open Source
  - TABARI
- Proprietary
  - VRA-Coder [VRA]
  - JABARI-NLP [Lockheed]
  - Event triples system [BBN]
  - Xenophon [SAE]
  - Profiler-Plus coder [Social Science Automation]

Dictionaries: CAMEO

- CAMEO Verbs
  - 15,000 verb phrases
  - Organized into WordNet categories [“soon”]
- CAMEO agents: also WordNet
- CAMEORCS: Religious classification system for 1500 religions
- Ethnic codes: roughly 600 ethnic groups

Dictionaries: ICEWS

- ICEWS International dictionaries
  - NGO/IGO
  - MNC
  - Militarized non-state actors
- ICEWS global state-level dictionaries
  - 160 countries
  - Generally based on frequency using NER software; includes individuals and political organizations
  - Stored in XML in custom format
  - May or may not be released

CountryInfo.txt

File contains about 32,000 lines, covering about 240 countries and administrative units (e.g. American Samoa, Christmas Island, Hong Kong, Greenland). It is internally documented and almost but not quite XML: The major fields are delimited with tags of the form <tag>...</tag> but elements inside are delimited with line feeds.

- Country name in English
- Adjectival forms and synonyms of the country name, including some non-English versions of the name
- ISO-3166 numeric, alpha2 and alpha3 codes, FIPS-10 code, IMF code, COW alpha and numeric codes
- Capital city
- Cities with populations over 1-million
- Regions and geographical features (WordNet meronyms)
- Leaders, 1960-2008 (rulers.org)
- Members of government, 2003-2010 (CIA World Leaders)

Named-entity Recognition/Resolution

- PoliNER/CodeCatcher: open source Python program which detects names based on capitalization patterns and sorts these by frequency. CodeCatcher is a machine-assisted coding utility for generating CAMEO-style dictionaries from this output.
- Numerous other NER programs exist as open-source; generating names from political texts is relatively easy
- Open issue: automatically generating multiple forms of a name based on the full version of the name provided in the CIA World Leaders
- Open issue: efficiently updating changes in status (death, removal from office, etc)

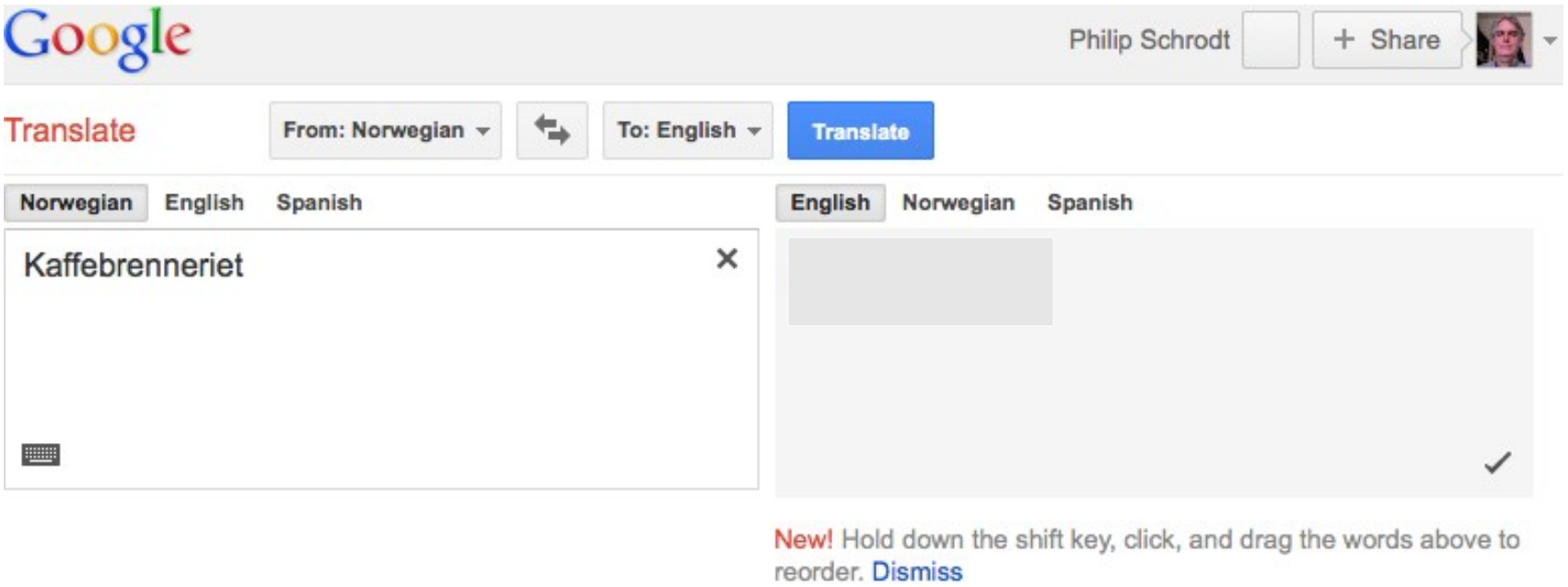
Additional open source NLP tools

- Sentence delimiting
- Pronoun and noun-phrase coreferencing
- Noun/verb disambiguation
- Full parsing
- Geolocation tagging
  - Decidedly mixed results in ICEWS experiments
- Machine translation

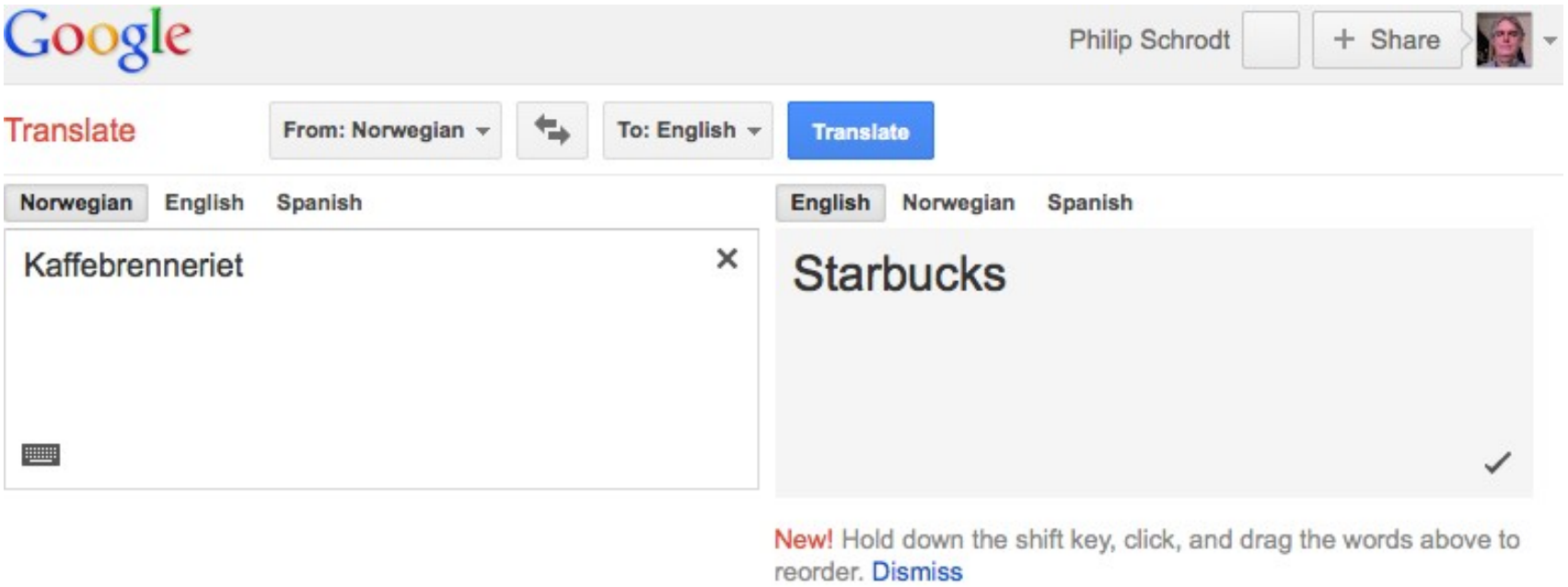
Additional open source NLP tools

- Sentence delimiting
- Pronoun and noun-phrase coreferencing
- Noun/verb disambiguation
- Full parsing
- Geolocation tagging
  - Decidedly mixed results in ICEWS experiments
- Machine translation
  - Maybe...

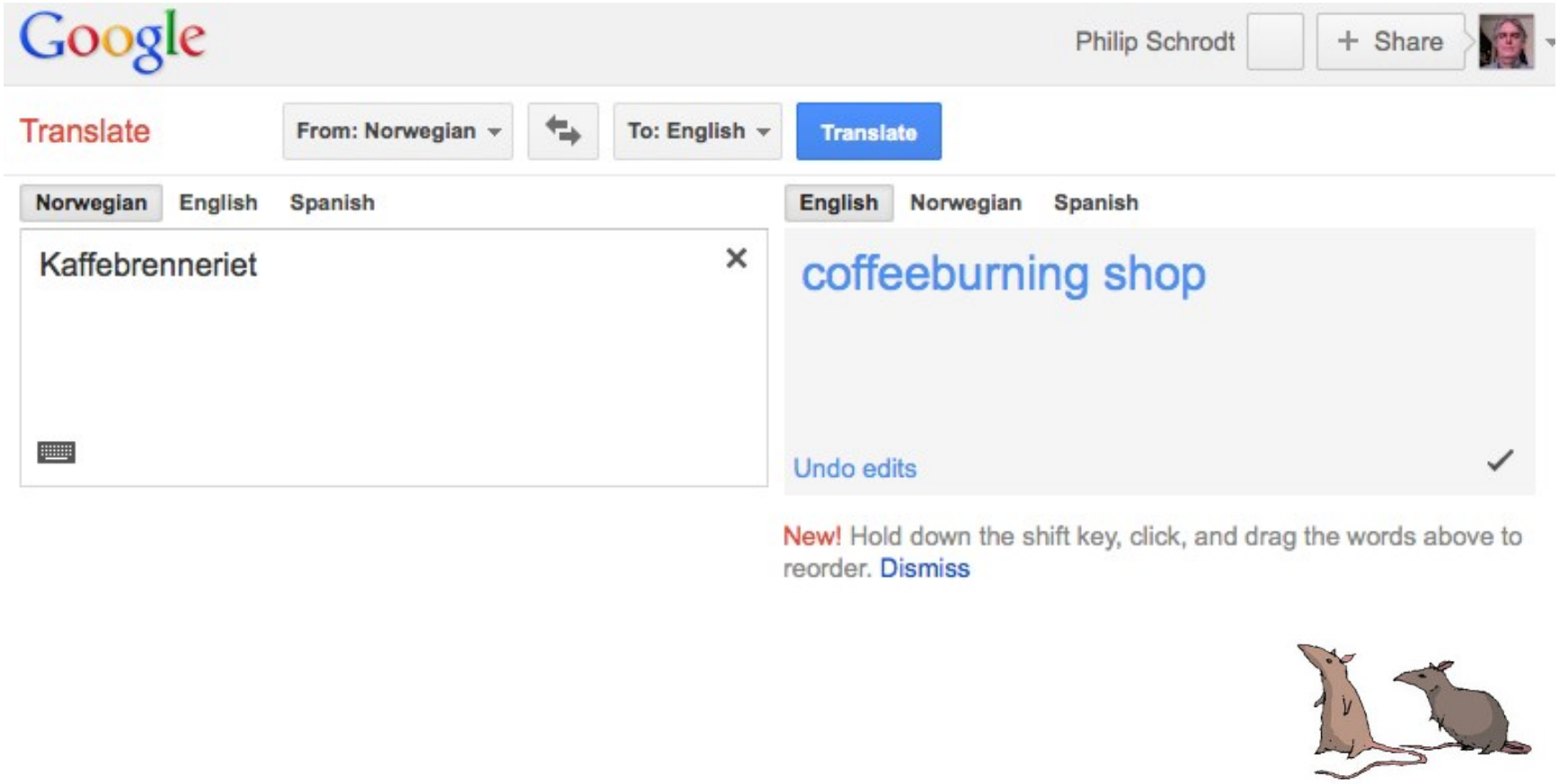
Google Translate in action....



Google Translate in action....



Or this alternative...



Near-real-time Text Sources

- RSS feeds from individual newswires and newspapers
- Google News
- European Media Monitor
- Open Source Center (CIA)
- Lexis-Nexis
  - Search engine remains dodgy
- Factiva
  - Reliable but expensive

Near-real-time Coding

- Integration of downloads, formatting and filtering is trivial in a Unix system
  - Coding with a lag of about 24 hours seems about right to allow corrections of reports and duplicate detection
- EMM produced about 10Gb of text per month, mostly HTML code which had to be filtered out
  - HTML formats vary dramatically between sources
- Formats do not remain constant and require some updating
- Limited archives

Machine-assisted Coding Tools

- Support vector machines for story filtering (MID, GTDS)
- TABARI as a pre-filter (Dugan and Chenoweth)
- SPEED has developed a suite of pre-processing tools

Questions?

Philip A. Schrodt

Political Science

Pennsylvania State University

University Park, PA 16801 USA

schrodt@psu.edu

http://eventdata.psu.edu