

# creating an automated event data system for arabic text

---

Andy Halterman, Jill Irvine, Christan Grant, Khaled Jabr, Yan Liang  
4 April 2018

- text is a key source of data for political scientists
- increasing use of automated text analysis
- most automated analysis is for English

Who did what to whom:

*“German Chancellor Angela Merkel criticized the Turkish government for its restrictions on free speech”*

Actor: German Chancellor Angela Merkel → DEU GOV

Target: Turkish government → TUR GOV

Event: “criticized... restrictions on speech” → CONDEMN

Using the CAMEO event ontology

Large ecosystem of tools for event data:

- Petrarch2 (event coder)  
<https://github.com/openeventdata/petrarch2>
- phoenix\_pipeline (end-to-end event data)  
[https://github.com/openeventdata/phoenix\\_pipeline](https://github.com/openeventdata/phoenix_pipeline)
- Birdcage (faster, distributed pipeline)  
<https://github.com/openeventdata/birdcage/>
- Mordecai (text geoparsing)  
<https://github.com/openeventdata/mordecai/>

*“German Chancellor Angela Merkel criticized the Turkish government for its restrictions on free speech”*

1. grammatical parsing of the sentence
2. find actor and event text
3. compare to dictionaries

Steps 1 and 2 are easy to change across languages. Step 3 requires unique dictionaries for each language.

The main bottleneck in making custom event data is in customizing dictionaries.

Understanding the best way to create dictionaries is useful for:

1. Making event data in other languages
2. Making dictionaries for new event types
3. Understanding where dictionaries come from

# making actor dictionaries (text approach)

Take me to fast actor coding

Take me to wkd coding

Sign Out    Track Your Performance

في مجال الشؤون العسكرية: اطلع المجلس على ما رفعه مجلس الدفاع المشترك بشأن مراحل تطوير قوات درع الجزيرة المشتركة وفقاً لقرارات المجلس الأعلى في دوراته السابقة، وكذلك مدى التنسيق والتعاون القائم بين دول المجلس في كل ما من شأنه تعزيز وتطوير الدفاع المشترك بين الدول الأعضاء، وبارك المجلس الأعلى ما تم إنجازه، ووجه باستكمال ما يتعلق به من خطوات وإجراءات في ...

**Nouns:** [مجال لشؤون، "التنسيق"، "الدول"، "تعزيز"، "شأن"، "تطوير"، "تعزيز"، "القائم"، "كل ما"، "الدفاع"، "خطوات"، "الأعضاء"، "تطوير الدفاع"، "مجلس"، "لشؤون"، "إجراءات"، "وفقاً"، "استكمال"، "دفع الجزيرة"، "دورات"، "المشتركة"، "الإنجاز"، "الأعلى"، "تطوير قوات"، "الدول الأعضاء"، "المجلس"، "الجزيرة"، "القرارات المجلس"، "هذا"، "السد"، "قوات درع"، "مراحل"، "التعاون"، "السابقة"، "شأن"]

**Verb:** [بارك، "يتعلق"، "رفع"، "وجه"، "تم"، "اطلع"]

- بارك
- يتعلق
- رفع

Search Sentences (Please input key words):

Go

Random Sentences    Commit

### Nouns Tagging Section

Role    Role-Summary

Actor Text     Target     Other    Add another role period

Actor Text:     **Synonym**    **CLEAR ALL**

Country:

Primary Role:

Secondary Role:

from:  ("yyyy-mm-dd")

to:  ("yyyy-mm-dd")

not sure?    **commit**

**Synonym List**

Verb tagging section

Verb:     **CLEAR ALL**

Verb Code:

# making actor dictionaries (ner approach)

### Fast Arabic Actor Coding (PERSON ENTITY)

Switch To Org Entity Total Left:103632

START COMMIT Skip

Country...

Primary...

Secondary...

Start Date:

End Date:

#### Related Sentences

- كما تكرر الشهيد ذاته في السودان التي عرفت إحدى أهم تجارب الانفتاح الديمقراطي في منتصف الثمانينات بعد الانتفاضة التي أطاحت حكم التميزي، فكانت «ثورة الإنقاذ» التي أجهضت التجربة في مهبها، بطن دموي عال ما يزال هذا البلد العربي الإفريقي المهم يتلغمه.
- وفي السودان، زجوا في السجن بأحد الأستاذة في عهد جعفر التميزي، فذهب عميد الجامعة لسيدة الرئيس بتوسط زميله.
- وأفاد التميزي أنه يتبع الفكر الأشعري وأن المجموعة التي ينتمي إليها «كانت بمسند إنشاء تنظيم شبيه بحزب الله لمحاربة إسرائيل وأنه عندما علم أن «أبو إسماعيل» من تنظيم «القاعدة» قطع عنه كلياً.
- وبعد رجوع التميزي من القاهرة بعد تشييع جنازة عبد الناصر، شاء أن يفتش إحدى المؤسسات.
- كما تردد الكثير من النكات عن عبد الناصر وصدام حسين ويورقية، نسج السودانيون ما يكفهم من النكات عن التميزي.



# making actor dictionaries (wiki approach)

Click the start button to get start.

Name:

Role:

Date Start:2008-10-1

Date End:Incumbent-01-01

Not Sure?

[wiki link](#)

# verbs (text approach)

Verb: ["استنر"]  
• استنر

Search Sentences (Please input key words):

**Administration Tool**

Verb:

Verb Code:   
 not sure?

- 000:NA
- 010:Make statement, not specified below
- 011:Decline comment
- 012:Make pessimistic comment
- 013:Make optimistic comment
- 014:Consider policy option
- 015:Acknowledge or claim responsibility
- 016:Reject accusation, deny responsibility
- 017:Express in symbols etc.

method	total actors coded	total verbs coded
regular interface	6,387	1,628
wiki translation	5,696	NA
NER coding	179 (with 6,667 skipped)	NA
wiki bio coding	2,327	NA
CAMEO translation	NA	~9,000

- Use Wikipedia where possible. Use raw text as a last resort
- Start by translating existing verb dictionaries, but understand limitations

- How different is data created from Arabic text from English text? What mistaken inferences might be drawn from relying on English language text?

- How different is data created from Arabic text from English text? What mistaken inferences might be drawn from relying on English language text?
- Pre-war political mobilization and violence against civilians in civil war (Balcells 2017). Do protests provide the same signal as election returns? Does the theory hold in Syria?