

Evolution of Data Needs in DARPA Language Programs

Bonnie Dorr

LDC Workshop on **The Future of Language Resources**

September 7, 2012





Data Needs in DARPA Language Programs

Past Successes:

- Data with general characteristics
- High resolution/quality, low noise
- Formal and semi-formal (mostly text, clean speech, machine printed)
- Standard dialects
- Lots of universal data (including parallel corpora)
- Slow acquisition and creation

New Challenges:

- Data with operational characteristics
- Low resolution/quality, high noise and degradation
- Informal genre and multiple modes (text, speech, handwriting)
- Multiple dialects
- A smaller amount of data – but of a more targeted nature
- Fast acquisition and creation (crowd-sourcing)



Examples

MADCAT:

- OCR'ed data: Program generated, now moving to real-world data
- Low noise in program-generated, extremely noisy real-world data.
- Large volumes

RATS:

- Clean speech transmitted through extremely noisy channels, not by mixing noise and signal
- Real-world data extremely noisy: 0-10db SNR
- Large volumes

GALE/BOLT:

- Initially formal and semi-formal (newswire, fora, news, talk shows)
- Lots of annotation with PropBank representations
- Moving to informal, dialectal, noisy data (emails, msgs, conversations)
- Dependency parsing, alignment, semantic roles

DEFT:

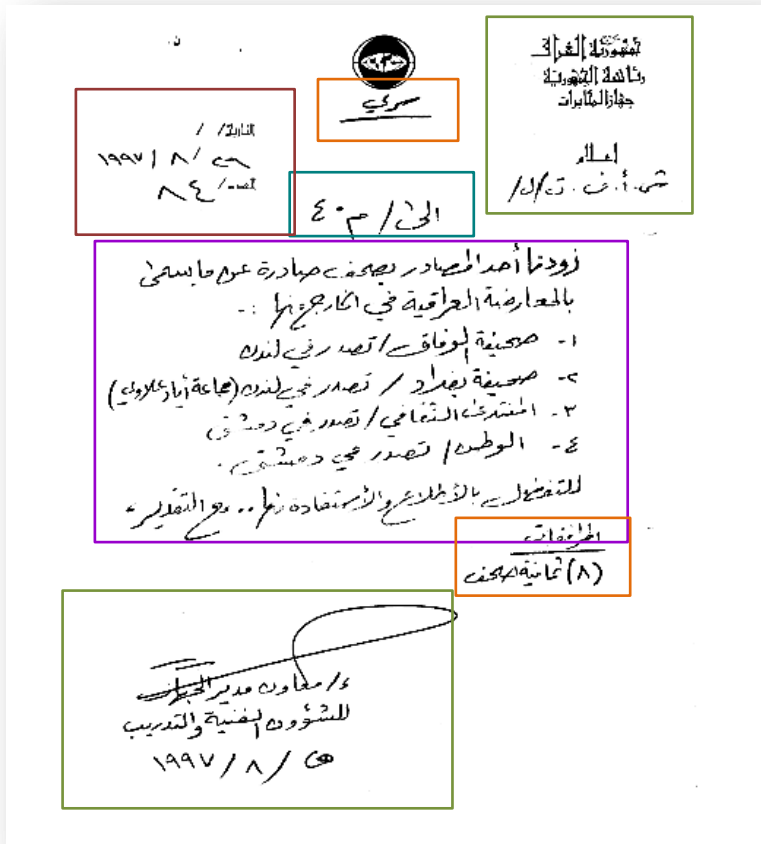
- Highly informal text and conversations
- Requirement for deep language understanding with richer annotations: entities, events, relationships, marking of implicit information.
- Training size small, volume high in real-world environment



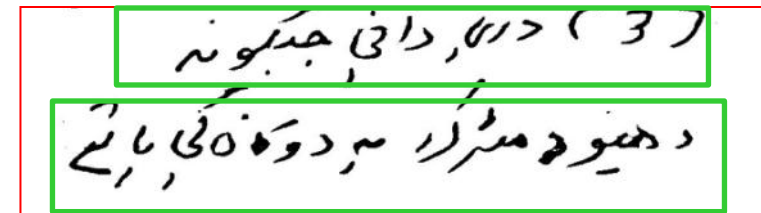
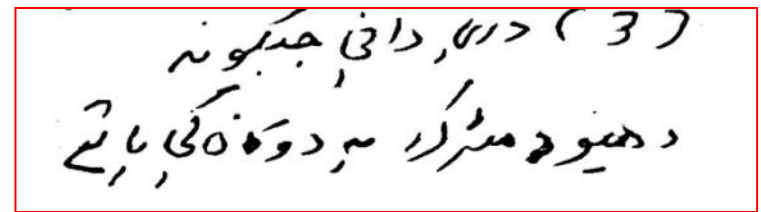
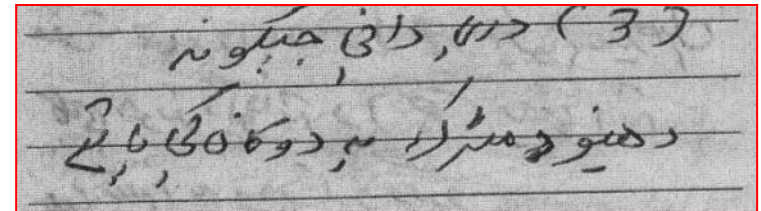
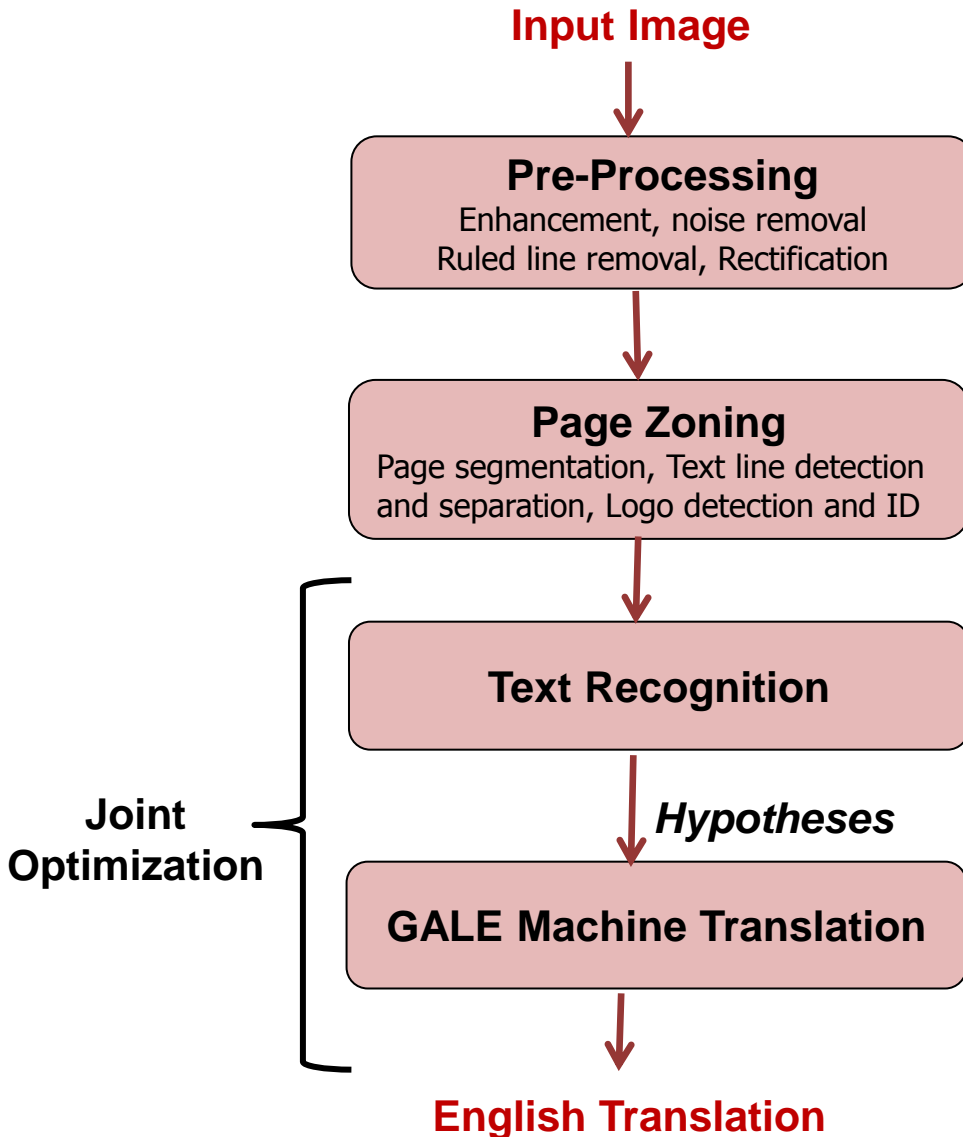
MADCAT Overview: Program Goal

Enable English Speaking Warfighters to Absorb & Analyze All Incoming Information in a Timely Manner

- Media: Hardcopy documents (machine-printed, hand-written)
- Language: Arabic
- Genres: Newspapers, magazines, letters/notes, ledgers



```
<xml>
<meta ID="BIAP-2003-000090.pdf"
  genre="correspondence" language="Arabic" >
<date obtained="Jun-10-2005" created="Aug-26-1997">
<author title="Assistant manager training and technical
  Department" address="The Republic of Iraq, President's
  Office, Intelligence Department">
<recipient name="m40">
<comment #1 translation="Confidential">
<text translation="One of the sources informed us of
  newspapers that are so called the Iraqi defiance abroad,
  some of them are:"
  <list><item 1-Al Wafaa newspaper published in
  London>
  <item 2-Baghdad newspaper published in London (Ayad
  Alawi group).>
  <item 3-Al Muftada al Thakafi published in Damascus>
  <item 4-Al Wattan published in Damascus></list>
  "for your information. Regards">
<comment#2 translation="Attachments 8 pages">
</xml>
```





MADCAT Program Collection:

- GALE Arabic text handwritten by ~350 authors
- Exhibits real-world variations in writing style, speed, and material
- Pages with and without rule lines
- High resolution, low noise
- 42K pages made available through Phase 3
 - No planned collections in P4 and P5

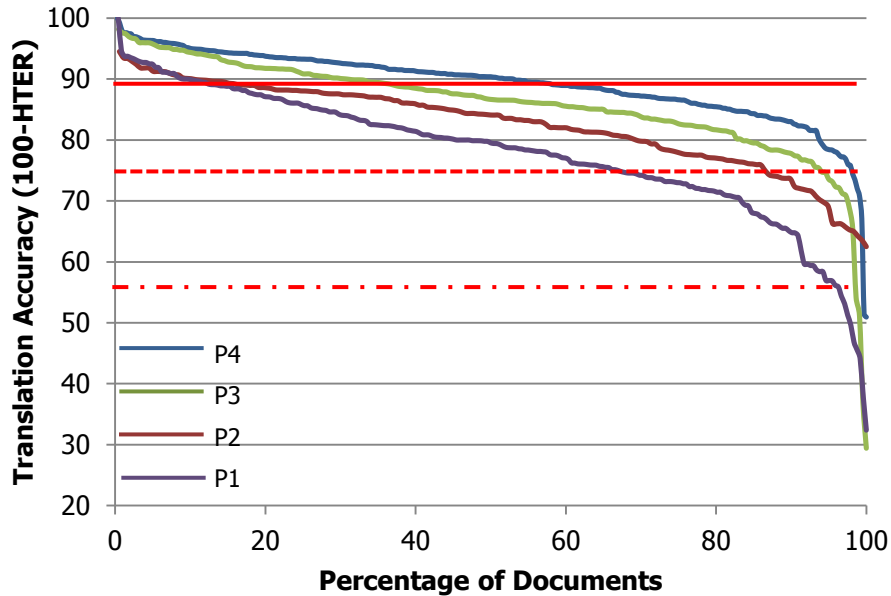
Anfal Collection:

- 18K pages of real-world documents collected in Iraq in 1991
- Poor scanning, low (unknown) resolution, varying degree of noise from low to severely degraded
- Mostly mixed type with significant variations in amount of handwritten (HW) vs. machine-print (MP) content
- Multiple genres (memos, letters, forms, tables)

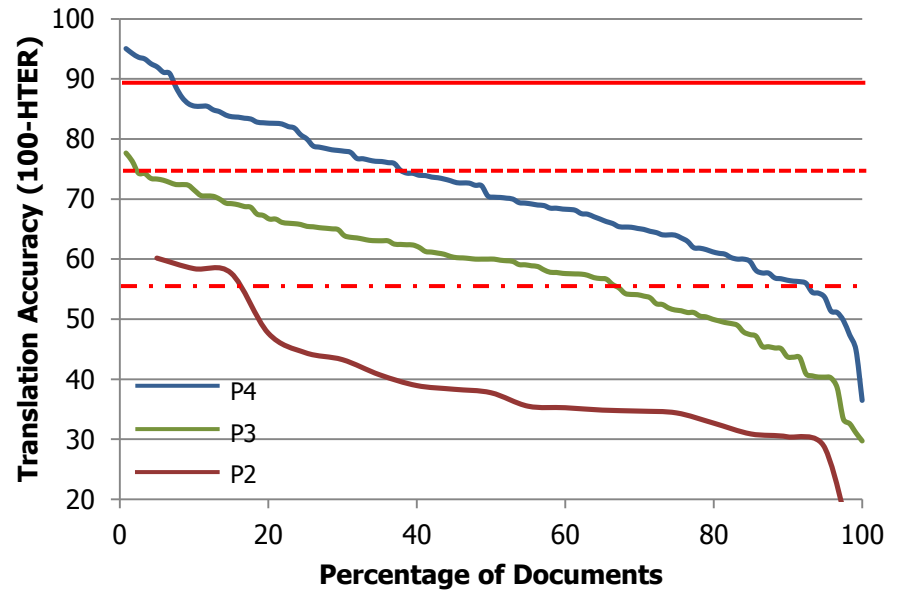


MADCAT Evaluation Results

MADCAT-Generated Data



Data collected in Iraq in 1991



- Editable
- - - Gistable
- . - Triageable



MADCAT Samples – LDC Data

نخزة 7 يونيو / شينخوا /
 ذكرت مصادر أمنية فلسطينية وشهود عيان ان
 احد نشطاء حركة فتح قتل صباح اليوم
 الخميس ١ واصيبت تسعة مواطنين اخرين في
 تجديد الاشتباكات بين عناصر من فتح وحماس
 في حي تل السلطان في مدينة رفح جنوب قطاع
 غزة.
 وقالت المصادر ان هؤلاء وانزل وصبى
 (27 عاما) من نشطاء حركة فتح قتل واصيب
 تسعة مواطنين بينهم ثلاثة اطفال في
 تجديد للاشتباكات المسلحة بين عناصر من
 فتح وحماس في منطقة تل السلطان في رفح
 جنوب القطاع مشيرة الى ان اثنين من
 المصابين في حالة خطيرة للغاية.
 وقال شهود عيان ان الاشتباكات عنيفة وقعت
 بين حركتي فتح وحماس منذ ساعات الصباح
 مؤكدة أنها ما زالت مستمرة حتى الان.
 وأُفاد شهود عيان أن مسلحي الحركتين
 المنشورا في شوارع الاعي.

ثلاث دول برلمانية تعترض على المرسوم
 الرئاسي بخصوص الجمعيات والهيئات
 اخذت ثلاث برلمانية ان المرسوم الذي
 اصدره الرئيس بشأن الجمعيات والمؤسسات
 الاصلية يشغل سابقا من صريح انه قد قام
 فعليا بإلغاء العمل بقانون الجمعيات دون
 أن يعلن ذلك، بالإضافة إلى توثيق اجراءات
 يتمسك الى حد بعيد بالحياة الديمقراطية في
 فلسطين، ويعد إلى درجة كبيرة من الحريات
 الفردية والجماعية.
 وقالت لجنة وقائمة الشهود أبو علي مصطفى
 وقائمة الجدل وقائمة فلسطين
 المستقلة - المبادرة الوطنية الفلسطينية
 في بيان وصل "معا" نسخة عن "ال
 الواضح ان هذا المرسوم يتعارض مع
 القوانين المعمول بها في حرة عمل
 الجمعيات والمؤسسات والهيئات ويعطي
 صلاحية مطلقة المصالح التلويحي بالتدخل في
 شؤونها دون مسوغ قانوني أو اجرائي
 عملي، حسب ما نص عليه قانون الهيئات
 والمؤسسات الأهلية رقم (١) لعام
 2000".

السيد محمد ابي المظفر

١٩٧٧ - ٥ - ٢٢

سيد: صبا اواركم يا تقي (بروز) بالمعنى سيد محمود) وفتح معه الموضوع ووافقت
 سيد محمود ما أمرتم به ما سئلكم رسالة شخصية مع أحد أفراد المفوضية المذكورة
 (محمود درويش) ويطلب منكم الجواب وينتظر أيا مركز معكمه نص رسالة على شكل
 رسالة رسمية:

السيد المحترم عزيزي من فضلكم والى غاية المثل: يوم الخميس ١٩٧٧ - ٥ - ١٩
 مع هذه الرسالة أقدم لكم قتيعة الحارة برحبها قبولها إلى (سيد محمود) من أهالي
 قرية (جده لندله) التابعة لنا حية ماوت منذ ٩ أشهر حملته السلاح مع جماعة كمال
 كركوكي وعلمنا ما تفرق مسؤوليات وبعثنا المباحة (أو من نصيب المقر العام) وفضلتون
 بالتوجه في كركوكي وقرأتكم هذا العرض أو من بعد المداولة واولا للاطلاع إرسال المداولة
 السيد عام عن اسراء جماعة البارئ على الميل مرة أخرى - جمع أعداد ميلادنا إلى أعداد
 من أهالي (شمال) وجاءوا إلى من نحننا ومن قبلنا ٣ سورا الميل برحبها قبولها حيدودنا
 من أنتاج علي عسكري بدرنا بالتوسيع وقرأنا الرسالة البسطا بالان واليون والعسور
 وكان ذرية من نجهولكم. مننا شخصيا فتحت المدن وضع لي بأن السلام من ميلادنا
 الطريق وفي يتسهم للمرة الثانية تجر من الميلاد. من ناحية أبعاد ناس من كركوكي
 والعملاء وبقية فتحت آخر نقطة من ذمنا إن أكون حيدودنا مخلصا لحماية ميلادي وعلى
 ايدنا في قوة العرب عامة كركوكي وافتقا وبعثنا الحيا التالف الوحيد مننا الرجاء
 ايدنا في قتيعة التالف التالف لآله أحمد حسن البكر والرعية صدام حسين. وهما اللذان
 هذا البسطا في طغورنا ايدنا اللذرية ايدنا سفيها ثروات ميلادنا وسفيها العراق. وبن الوشا
 التالف تعلقنا اعادة السلام ورضنا نية اله الموات الخبيث بالتحليل كردستان
 ايدنا بالسلام بعد ايدنا ايدنا علمت يا نيكنت ملك الطريق الاقرب ما سمعت شغري الى
 والى المقاتلة العسكرية في منطقتنا حيث أدت ايدنا حال مع سيداتكم بالتعاون مع وتوخ
 (محمد معروف) محمد درويش (كركوكي) وفي هذا التالف وسلم نفسنا على
 ثقة ايدنا في حالنا وعلمنا وعلمكم. وبيدنا من الحارين الحيا اللذرية سلكتنا سائقا
 ومن اليريدنا من بعدنا في جميع المواقف أنه يكون حيدودنا مخلصا من كركوكي من يريد
 تنجيب ميلادنا الحبيبة ونقسم ان لا نساك مرة أخرى من الحيا نية. وكوننا حيدودنا
 لجميع أواركم. جمع أواركم أو سلم نفسنا بعد نسين الوقت والمكننا من قتيعة
 حيدودنا ما سمعت سأ بقتي هذه الميثاقية على حياكم ومصالحنا الدولة ويقرر المستطاع
 أو يسلناكم تحركات وعملياتنا عن كركوكي وعلمنا عسكري
 سيد: إننا تفرقون علمنا الينا مع المخرجه من يتجلى علمنا التالف التالف:

السيد محمد ابي المظفر

١٩٨١ - ٥ - ٢٢

السيد: صبا اواركم يا تقي (بروز) بالمعنى سيد محمود) وفتح معه الموضوع ووافقت
 سيد محمود ما أمرتم به ما سئلكم رسالة شخصية مع أحد أفراد المفوضية المذكورة
 (محمود درويش) ويطلب منكم الجواب وينتظر أيا مركز معكمه نص رسالة على شكل
 رسالة رسمية:

السيد المحترم عزيزي من فضلكم والى غاية المثل: يوم الخميس ١٩٨١ - ٥ - ١٩
 مع هذه الرسالة أقدم لكم قتيعة الحارة برحبها قبولها إلى (سيد محمود) من أهالي
 قرية (جده لندله) التابعة لنا حية ماوت منذ ٩ أشهر حملته السلاح مع جماعة كمال
 كركوكي وعلمنا ما تفرق مسؤوليات وبعثنا المباحة (أو من نصيب المقر العام) وفضلتون
 بالتوجه في كركوكي وقرأتكم هذا العرض أو من بعد المداولة واولا للاطلاع إرسال المداولة
 السيد عام عن اسراء جماعة البارئ على الميل مرة أخرى - جمع أعداد ميلادنا إلى أعداد
 من أهالي (شمال) وجاءوا إلى من نحننا ومن قبلنا ٣ سورا الميل برحبها قبولها حيدودنا
 من أنتاج علي عسكري بدرنا بالتوسيع وقرأنا الرسالة البسطا بالان واليون والعسور
 وكان ذرية من نجهولكم. مننا شخصيا فتحت المدن وضع لي بأن السلام من ميلادنا
 الطريق وفي يتسهم للمرة الثانية تجر من الميلاد. من ناحية أبعاد ناس من كركوكي
 والعملاء وبقية فتحت آخر نقطة من ذمنا إن أكون حيدودنا مخلصا لحماية ميلادي وعلى
 ايدنا في قوة العرب عامة كركوكي وافتقا وبعثنا الحيا التالف الوحيد مننا الرجاء
 ايدنا في قتيعة التالف التالف لآله أحمد حسن البكر والرعية صدام حسين. وهما اللذان
 هذا البسطا في طغورنا ايدنا اللذرية ايدنا سفيها ثروات ميلادنا وسفيها العراق. وبن الوشا
 التالف تعلقنا اعادة السلام ورضنا نية اله الموات الخبيث بالتحليل كردستان
 ايدنا بالسلام بعد ايدنا ايدنا علمت يا نيكنت ملك الطريق الاقرب ما سمعت شغري الى
 والى المقاتلة العسكرية في منطقتنا حيث أدت ايدنا حال مع سيداتكم بالتعاون مع وتوخ
 (محمد معروف) محمد درويش (كركوكي) وفي هذا التالف وسلم نفسنا على
 ثقة ايدنا في حالنا وعلمنا وعلمكم. وبيدنا من الحارين الحيا اللذرية سلكتنا سائقا
 ومن اليريدنا من بعدنا في جميع المواقف أنه يكون حيدودنا مخلصا من كركوكي من يريد
 تنجيب ميلادنا الحبيبة ونقسم ان لا نساك مرة أخرى من الحيا نية. وكوننا حيدودنا
 لجميع أواركم. جمع أواركم أو سلم نفسنا بعد نسين الوقت والمكننا من قتيعة
 حيدودنا ما سمعت سأ بقتي هذه الميثاقية على حياكم ومصالحنا الدولة ويقرر المستطاع
 أو يسلناكم تحركات وعملياتنا عن كركوكي وعلمنا عسكري
 سيد: إننا تفرقون علمنا الينا مع المخرجه من يتجلى علمنا التالف التالف:



Translation of Program Generated Data

Gold Standard

Gaza, June 7, Xinhua:

Palestinian security sources and eyewitnesses said that a Fatah activist was killed this morning, Thursday, and nine other citizens were wounded in renewed clashes between elements of Fatah and Hamas in the Tal al-Sultan neighborhood, in the southern Gaza Strip city of Rafah.

The sources said that Fatah activist Fuad Wa'il Wahbi (27 years old) was killed and nine citizens, including three children, were wounded in renewed armed clashes between elements of Fatah and Hamas in the Tal al-Sultan area in Rafah in the southern part of the Strip. They noted that two of the wounded are in extremely serious condition.

Eyewitnesses said that fierce clashes have taken place between Fatah and Hamas movements since the morning hours, and they confirmed that it is still continuing.

Eyewitnesses said gunmen of the two movements were deployed in the neighborhood streets.

Machine Translation

GAZA, June 7 (Xinhua).

Palestinian security sources and eyewitnesses reported that the activists of Fatah movement, was killed today, Thursday morning and other nine citizens were injured in renewed clashes between Fatah and Hamas elements in Tel El-Sultan district in Rafah city, south of Gaza Strip. The sources said that Fouad Wael (27 years old) from the Fatah Movement activists killed and nine citizens were injured, including three children, in armed clashes renewed between members of Fatah and Hamas in Tel El-Sultan area in Rafah, south of Gaza Strip, pointing out that two of the injured are in a serious condition.

Eyewitnesses said that violent clashes took place between the Fatah and Hamas movements since the early morning hours, confirming that it is still continuous till now.

Eyewitnesses reported that the Muslims of the two movements were deployed in the streets of the district.

Accuracy = 92%



Translation of Field-Collected Data

Machine Translation

Mr. respected security director.
In principle, according to remind you called (banned will claim) opened the subject with him.
Mr. Mahmoud ordered agreed.
We sent you a personal message with one of the group members of the center, Darwish) and to you the answer.
It is expected that.
This is the text of his message of thought.
The text of the message. "
Therefore, that will be headed by the family and secular voters on Thursday the name ".
With this letter to you his warm greetings to accept it.
It will support from the people of the village of (his) bout.
9 months ago a campaign with Group Kamal my company and i responsibilities.
Will be sold and a faction of the general headquarters of the year).
He says that in the Central.
And offer their lives since the first show after the attempt and bacteria.
Year after the collapse of the customer Barzani once again a quarter of the enemies lead us to prepare for its withdrawal from the side of the Santana Iran and not to our rules of the organization and the the (metal).
In Canada, the to the region and the Syrian regime.
Reem agents without and followers on military to sabotage and threatened the simple people in the Iraqi issue and attacked them.
Personally, I am now explain to me that all foreigners and agents and for the second time in their intention to sabotage the country.

Accuracy= 62%

Machine Translation

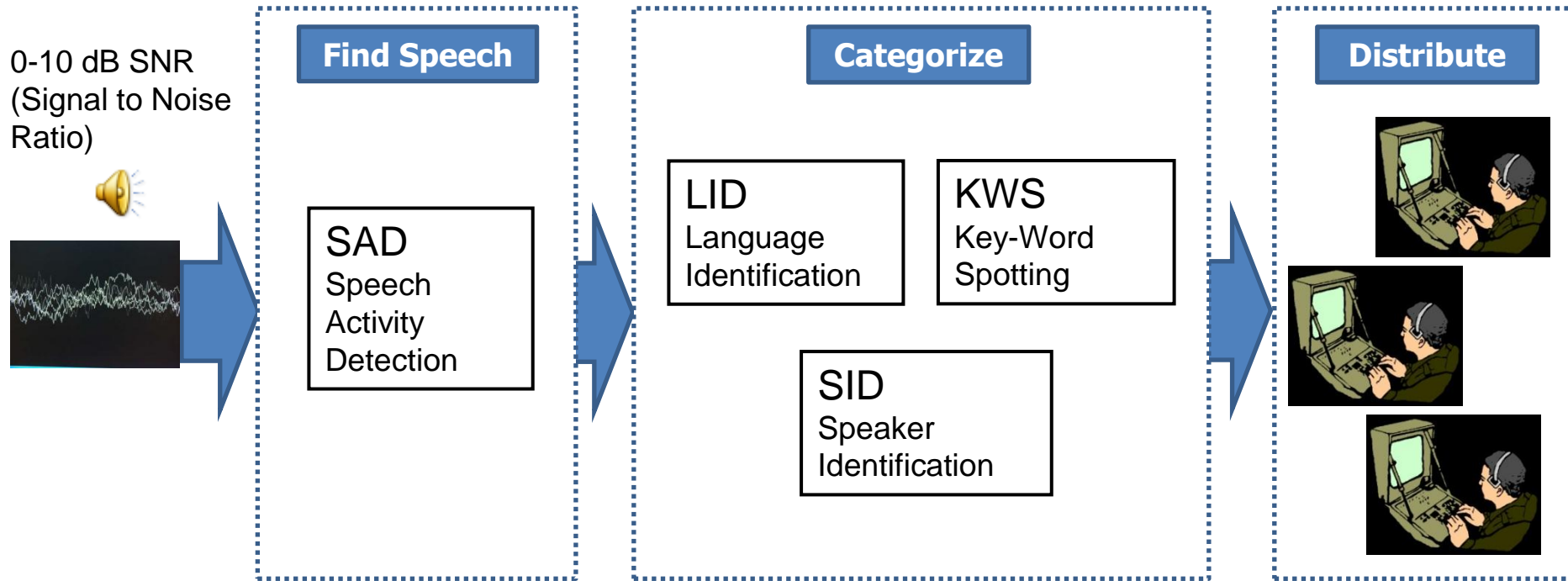
Them.
, the company.
_ 2 in his team Khartoum or not. "
Manhattan Project.
Anger.
, i. "
Baker says the resolution sets minorities in my village of military, I mean, we as of March/19.
Daily research group escaped from the military service.
Thailand.
He accused, I mean, including a leader in "I want to.
Alert service five Monday, Strategic Studies of the Kurds doubt in lists salaries will be open for stones is accompanied by a son in Indonesia, and then spread and stability as the colony warned North/1980 better. Start tomorrow. You have experience in the euro, Yemeni Minister of my life.

Accuracy= 34%



RATS: Program Goal

Create technologies for exploitation of potentially foreign speech-containing signals received over extremely noisy and/or highly distorted communication channels



Improve Capability to Find and Make use of Foreign Language Speech Signals



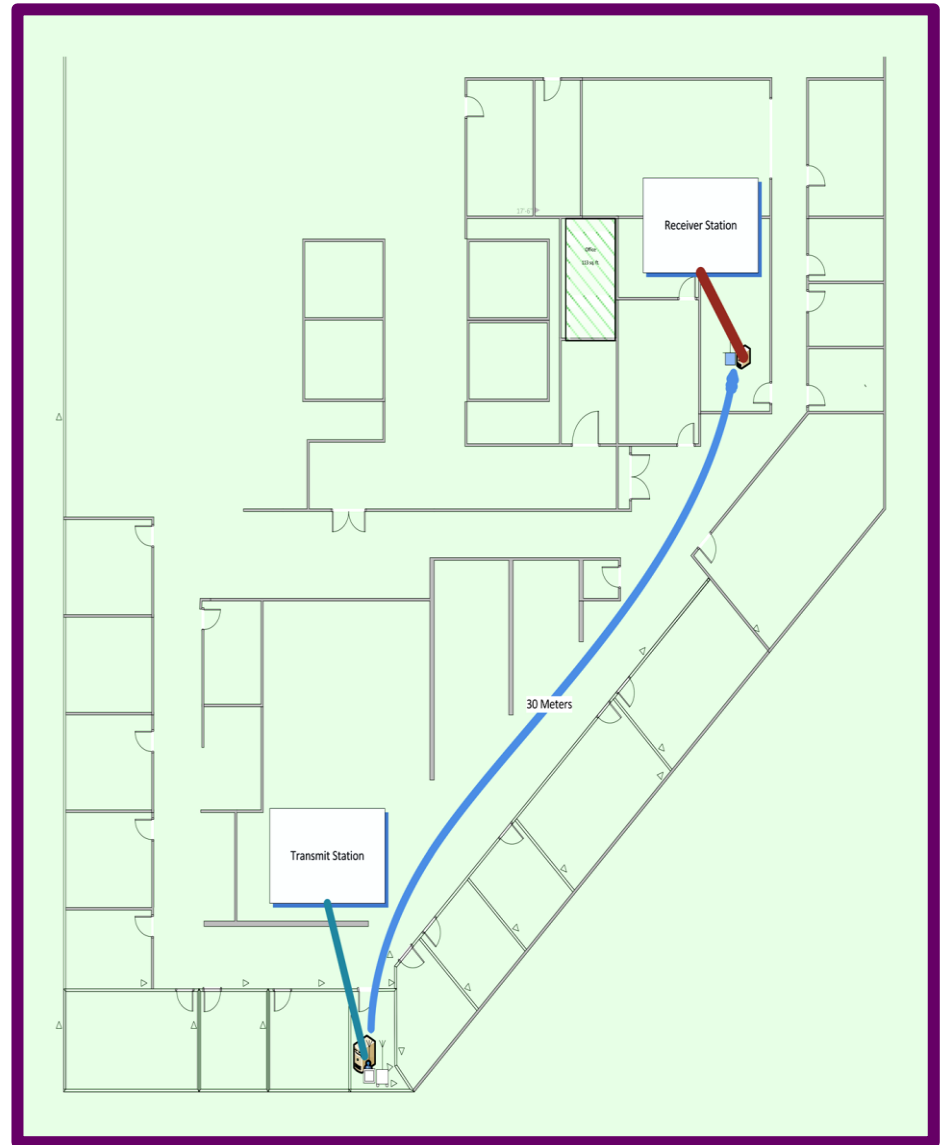
RATS Data

- RATS data consist of both program-generated data and field-collected data
- Program-generated data
 - Simulation of real conditions
 - Degradation is not achieved by a simple method of mixing noise
 - Speech is generated with degraded transmission facilities
 - Speech is generated in real-time or pre-recorded speech is re-transmitted
- Field-collected data
 - The program has secured some field-collected data
 - Will be used for evaluation
 - The evaluator has a classified testing facility
 - Could also be used for training
 - All fielded systems must be field trainable






















Transmission Path

- The Comms-Link collection systems are Located in two separate rooms
- The rooms are have a line-of-sight distance of approximately 30 meters
- However, there are significant structural obstacles between the two spaces
- This increase path loss for all system transceivers, and may produce multipath interference for shorter wavelength transceivers.





Degraded Speech Samples

RATS Channel	Intelligible		Unintelligible	
	English	Non-English	English	Non-English
A				
B				
C				
D				
E				
F				
G			None	
H				



In-The-Field Training Requirements

Technology		Amount	comments
General		Massive	Un-annotated Data
SAD		2h	1h speech, 1h noise. > 5 speakers
LID	Adapt to new channel	1h	Labeled, >20 sec per segment, >12 speakers
	Develop a new language	3h	Labeled, >20 sec per segment, multiple speakers and channels
SID	New Speaker Known Language and Channel	3 min.	6 samples, 30 sec. / sample
	Develop a new Channel	1h	20 speakers, 6 samples, 30 sec./sample multiple channels
KWS	New channel – Known language	1h	Fully transcribed
	Develop a new language	50h	Transcribed, a variety of channels



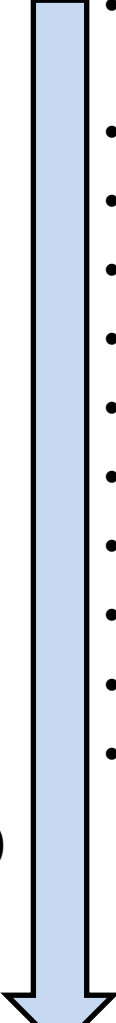
- **Genre staging**
 - Email (fora) → Messaging → Conversations
- **Language varieties**
 - English, Mandarin Chinese and Egyptian Arabic
- **Data volume targets**
 - 2Mw source data collected per language/genre
 - 10% of all collected data set aside for evaluation
 - 800Kw (40%) translated into English
 - 200Kw (10%) Egyptian data also translated to MSA
 - 400Kw (20%) annotated for Word Alignment, TreeBanking, PropBanking, Coreference



- Genre Staging
 - Narratives → Conversational speech/text → Foreign conversational speech/text
 - Implicit information essential
- Language varieties: yes
- Data volume targets
 - Approximately 1.2Mw of text per language/genre
 - Approximately 150 hours conversational per language
 - Annotated for Entities, Events, Relations, Coreference, and Regions of interest

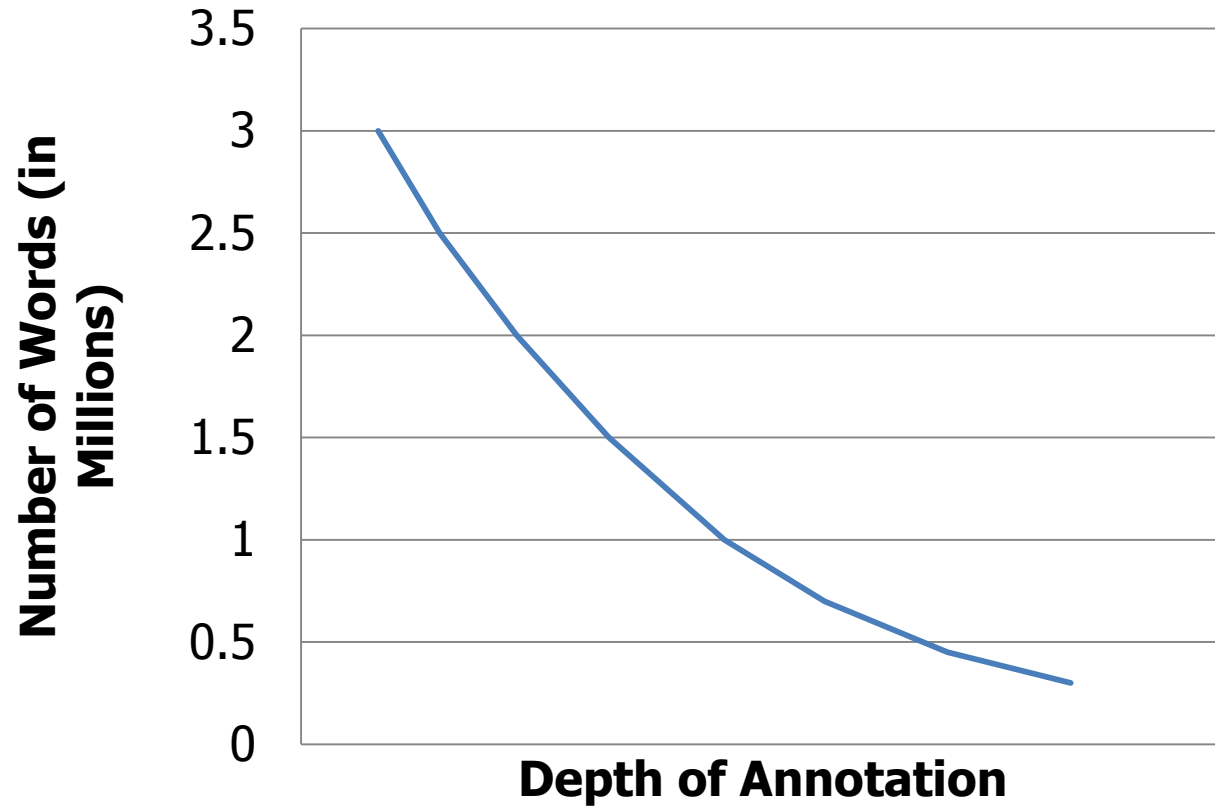


Depth Perspective

- POS-Tagged Brown and BNC
 - Penn Treebank (PTB)
 - WordNet
 - Sense Annotated SEMCOR
 - Prague Dependency Treebank
 - ACE Annotated Corpora
 - NomBank
 - TimeBank
 - PropBank
 - VerbNet
 - LCS-Tagged Corpora
 - FrameNet
 - Penn Discourse Treebank (PDTB)
 - Opinion-Tagged Corpora (MPQA)
- 
- Acoustic-Phonetic Continuous Speech Corpus (TIMIT)
 - OCR'ed Noisy Corpora (MADCAT)
 - Field-Collected OCR'ed Data (ANFAL)
 - Broadcast News
 - Conversational Telephone
 - Rich Transcription Corpora
 - Broadcast Conversations (Talk shows)
 - Dialectal Broadcast Collections
 - Noisy Transcript Corpora (RATS)
 - Two way communications (BOLT-B)
 - Other informal Communications
 - Forums
 - Emails
 - Chat
 - Messaging



Quantity-Quality Inverse Relation: How is this Possible?





Questions?

bonnie.dorr@arpa.mil