Resources for Arabic Natural Language Processing

Mohamed Maamouri, Christopher Cieri
{maamouri,ccieri}@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and
Department of Linguistics
www.ldc.upenn.edu

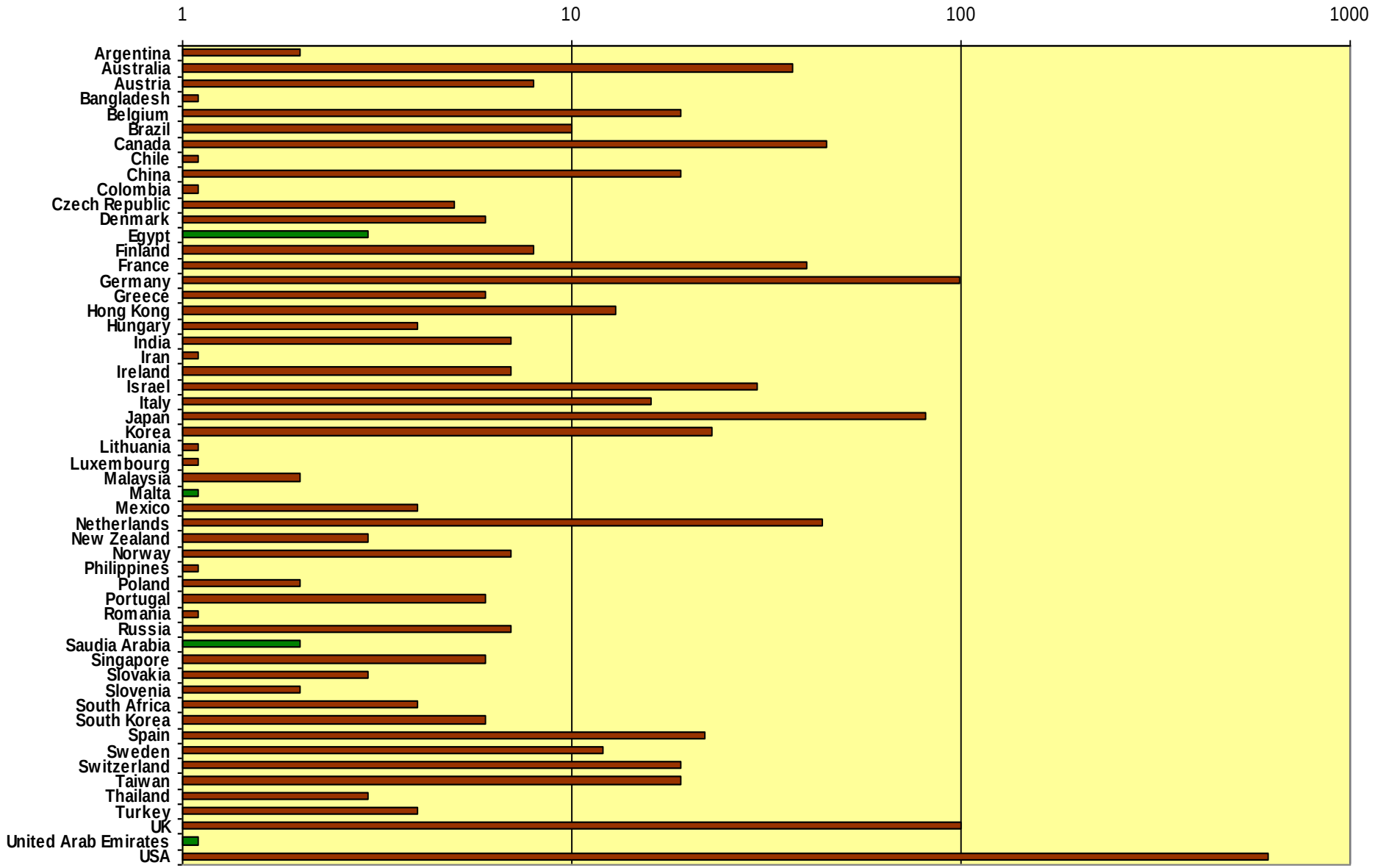
Background

- Language resources necessary component to language development
- Language resources expensive to create
  - require special skills/staff, specialized equipment
- Organizations that create language resources may not distribute
  - no interest, no infrastructure, reduce competitive advantage
- Problem: Lack of adequate supply of resources stands as an impediment to language development.
- Solution: Build non-profit language resource center to promote language development through the sharing of resources
- Acquire specialized equipment, develop specialized staff
- Build relationships with corpus authors, other data providers, and research communities
- Maintain permanent data archives with bug reports, re-releases, on-going rights to data
- Provide standard reference data to evaluate competing algorithms/analyses

LDC Roles

- Founded April 15, 1992 as a non-profit activity of the University of Pennsylvania
- Specialized publisher (>15,000 copies of 209 publications)
  - resources for linguistic education research and technology development
  - activities supported primarily through membership fees
- Open consortium: any organization interested in language resources may join (almost 1400 users)
- Intellectual property intermediary: negotiate agreements between data providers and data users
- Corpus Creator: create and annotate language resources to specification that can be widely shared
  - community initiatives, corporate and government sponsored projects, joint projects
- Research Group:
  - research approach to language resources
  - conduct research on standards and best practices

LDC Users Worldwide



Resources by Language

	Speech / Transcripts					
Language	*Broadcast*	*Telephone*	*WideBand*	*Parallel Text*	*Newswire/ Other Text*	*Lexicon*
Arabic (Egyptian)		■			■	■
Czech	■			■	■	
Dutch					■	■
English	■	■	■	■	■	■
French	■	■		■	■	
German	■	■		■	■	■
Hindi	■	■			■	
Japanese		■	■		■	■
Korean	■	■			■	■
Mandarin	■	■		■	■	■
Persian	■	■			■	■
Portuguese	■				■	
Russian	■	■			■	■
Serbo-Croatian	■				■	
Spanish	■	■	■	■	■	■
Tamil		■			■	
Thai	■				■	
Turkish	■				■	
Vietnamese	■	■				

Albanian, Arabic, Armenian, Azerbaijani, Bangla, Belorussian, Bosnian, Bulgarian, Burmese, Cantonese, Croatian, Czech, Dari, English, Estonian, Farsi, French, Georgian, German, Greek, Hausa, Hindi, Indonesian, Kazakh, Khmer, Kinyarwanda/ Kirundi, Korean, Kosovian, Kurdish, Kyrghiz, Laotian, Latvian, Lithuanian, Macedonian, Mandarin, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Tajik, Tatar-Bashkir, Thai, Tibetan, Turkish, Turkmen, Ukrainian, Urdu, Uyghur, Uzbek, Vietnamese

Coordinated Resources

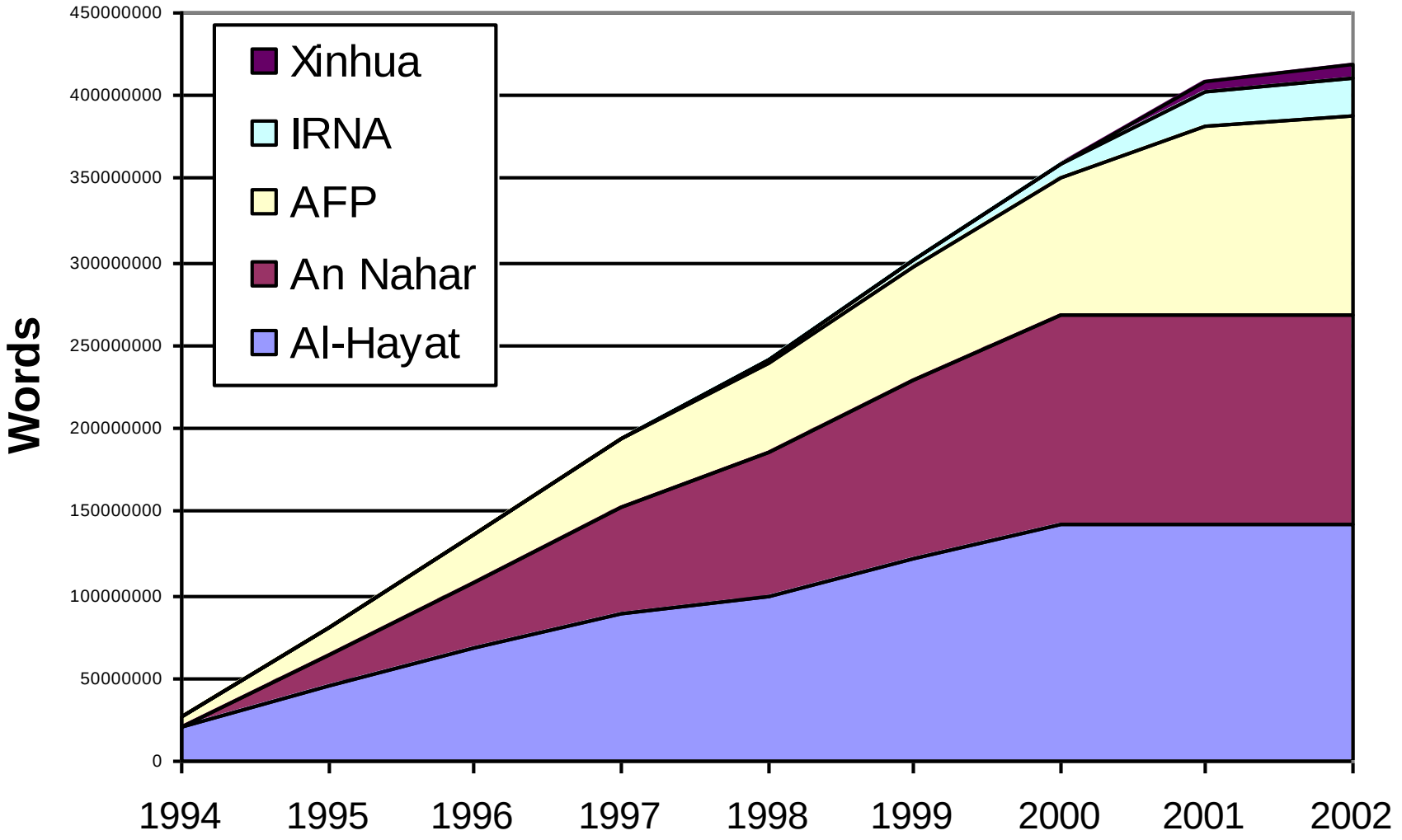
- Focus on major languages: English, Chinese, Arabic, Spanish
- Battery of Resources to meet major research and development needs
- Supporting: language modeling, speech recognition, translation, translingual information retrieval, natural language processing
- Resources also useful for any empirical language study including linguistic analysis, language teaching
- Gigaword News Text Corpora – 1B words, variety of news sources
- Parallel Text – pairs of documents and aligned translations
- Broadcast News – with time-aligned transcripts, important domain for its inherent interest and for its broad vocabulary
- Conversational Speech – telephone conversations and meetings, with time-aligned transcripts
- Pronunciation/Multilingual Lexicons – relate source word forms to:
  - set of target glosses, syntactic and frequency information, pronunciation, morphological analysis, optionally mediated through morphological analysis/synthesis engine
- Treebanks – text annotated to show the morpho-syntactic properties of sentences and their constituents
- Technology-Specific Evaluation Resources – MT & IR

Very Large Text Corpora

- Collecting news text since 1994
- Published Arabic Newswire, 76Mwords, in 2001
- To support robust modeling of rare phenomena need Gigaword News Text Corpora
- English, Chinese and Arabic
- Arabic: 480,000,000 words from Al Hayat, An Nahar, AFP, Xinhua, IRNA - looking for more
- Consistent encoding
- Light XML markup inline
- Other annotations should be stand-off

<DOC>
<DOCNO>19970304_AFP_ARB.0042</DOCNO>
<BODY>
<HEADLINE>
.. حملة واسعة لمكافحة
</HEADLINE>
<TEXT>
<P>
القاهرة - وقال المسؤول المصري ان التوصل الى لقاح ضد الطفيلي سيكون سابقة على درجة كبيرة من الاهمية لا سيما وان البلهارسيا تعتبر من بين الامراض الطفيلية الستة الاولى التي تضعها منظمة الصحة العالمية في اولويات برامج المكافحة الى جانب الملاريا ومرض عمى الانهار والجذام ومرض الفيل .
</P><P>
وينتشر في مصر نوعان من البلهارسيا، المعوية التي تشكل ثلثي الاصابات، والبولية . وينجم المرض عن دودة تعيش في المياه الراكدة وتكمل دورتها داخل القواقع ثم تنتقل الى الانسان عن طريق الجلد وتتوطن في الاوعية الدموية . ويسبب المرض في مراحله المتقدمة تشمعا في الكبد او فشلا كلويا وورما في المثانة . ويساعد الدواء المستخدم على العلاج من المرض لكنه لا يمنع ظهوره مرة ثانية .
</P><P>
وتنتشر البلهارسيا البولية (شيستوزوما هيماتوبيوم) بشكل خاص في شرق المتوسط وافريقيا، والمعوية (شيستوزوما مانسوني) وهي الاكثر انتشارا، في المناطق نفسها وكذلك في اميركا الجنوبية .
</P><P>
ويهدد المرض 600 مليون انسان خصوصا من سكان ضفاف الانهر والبحيرات، لا سيما في غياب انظمة حديثة للري والصرف الصحي . وتفيد احصاءات منظمة الصحة العالمية انه يتسبب بوفاة 200 الف شخص سنويا .
</P>
</TEXT>
</BODY>
</DOC>

Arabic Text Archive



Broadcast News

- **Goal: Database of broadcast news from around the Arabic speaking world, accurately transcribed**
- **Current 120 hours of Voice of America radio; 60 hours of Nile TV via SCOLA**
- **Topic Detection and Tracking Corpus – Phase 4 will contain Arabic broadcast news.**
- **40 hours will be carefully transcribed and released jointly with ELRA under the NSF-EU funded Networking Data Centers project**
- **Building capacity to collect additional source locally; also interested in partnerships.**

Conversational Arabic

- 1995 began collecting conversations in 18 linguistic varieties to support research in language identification and automatic transcription
- Included >450 telephone calls in Egyptian Colloquial Arabic
- 10 minutes from each of 200 calls transcribed, 120 of those released
- Publications include plain audio, time-aligned transcripts and a pronouncing lexicon
- Lexicon: surface form, romanization, pronunciation, morphological analysis and frequency in 3 data sets

249.44 251.14 B1:
و انتى عاملة إيه فى الدراسة بتاعتك

251.07 254.11 A:
والله الحمد لله قربت اخلص كتابة الرسالة

254.41 254.89 B1:
اه!

254.87 257.58 A:
و ربنا يسهل بقى ان شاء الله اناقش قريب يعنى

257.92 259.58 B1:
طيب ما ان شاء الله مبروك

259.45 261.03 A:
يا رب إدعى لى يا بابا والله

260.96 262.79 B1:
لا ربنا معاكى ان شاء الله

261.81 265.76 A:
اه يوفق!

TDT Corpora

- **Topic Detection and Tracking Corpora – support development of news understanding system**
  - **convert speech to text and segment into stories**
  - **identify new topics in the news and find all stories discussing a selected topic**
- **TDT-2 and TDT-3 together contain >1000 hours audio, >100K stories, annotated for relevance to 220 topics in English and Chinese**
- **TDT-4 will add 200 hours of Arabic news audio and transcripts plus newswire totaling 12,000 stories to similar amounts of English and Chinese annotated for 60 new topics**

30002.

Hurricane Mitch Chinese

Seminal Event

WHAT: Hurricane Mitch forms over warm ocean waters, killing thousands and causing millions of dollars in damage.
WHERE: The Caribbean and surrounding areas, particularly Honduras, Nicaragua and Central America.
WHEN: Mitch forms in late September 1998, and lasts through the month of October.

Topic Explication

Hurricane Mitch was the most destructive Atlantic hurricane since 1780, killing over 10,000 people in Central America and leaving millions homeless. **On topic:** coverage of the disaster itself; estimates of damage and reports of loss of life; relief efforts by the Red Cross and other aid organizations; impact of the hurricane on the economies of the effected countries.

Rule of Interpretation Rule 4: Natural Disasters

TREC CLIR

- **Text REtrieval Conference**
  - **organized by NIST, multiple tracks including SDR, CLIR, Q&A**
  - **broader topics than TDT, assessment replaces annotation**
- **CLIR 2001 Corpus is LDC Arabic New Corpus, 384,000 stories from Agence France Presse 1994-2000 and 25 topics; CLIR 2002 will add 50 topics**

Title	YES	NO	Total
Performing arts and Islamic institutions	383	471	854
Arab and western cinema	315	548	863
Traditional crafts and technology	133	898	1031
Arab cities and advertising pollution	88	1266	1354
Polio eradication in the Middle East	57	825	882
Measles immunization campaigns in the Middle East	17	645	662
Bilharzia/Schistosomaisis prevention in Egypt	24	949	973
Environmental protection laws in Egypt	57	668	725
Egyptian-Libyan relations during the 1990s	321	703	1024
Tourism in Cairo	242	683	925
Dead Sea archaeological finds	13	866	879
Information technology & the Arab world	132	958	1090
Water resources in the Nile Valley	100	664	764
Totals	**4122**	**18622**	**22744**

MT Corpora

- MT research lacks a stable metric to evaluate systems
- To support development of a metric, LDC created Multiple Translation Corpora
- >20,000 words, >100 stories in Chinese (Xinhua, Zaobao, VoA), Arabic (AFP, Xinhua)
- Selected from newswire and news broadcast to represent the mode story lengths
- Each story translated by at least 10 human translators, at least 3 systems to represent the range of translation practices and quality
- Translations are sentence aligned to original.
- Translations subsequently assessed by human judges
- Fluency – is the translation grammatical in target language?
- Adequacy – does story convey all information conveyed by idea translation?
- Chinese translations published in 2002; assessments to be added
- Arabic will be published with assessments in 2002.

Conclusions

- **We're keeping busy!**
- **This was just a survey of some resources; Maamouri will focus on part-of-speech tagged text and Treebanks**
- **So why is he here?**
- **Networks of coordinated resources are the way of the future**
- **Learning more about Arabic Processing**
- **Looking for additional resources**
- **Looking for users**
- **Looking for annotators**
- **Looking for collaboration that produces concrete results.**