# After the founding
# WHAT, WHY, HOW

Mona T. Diab
Columbia University

# One possible vision

## Sustainability via Science &Technology

# Basic Premise

- Goal of the UAE Arabic Project
  - Preservation of the Arabic language

# My assumption

- One of the goals is to create global citizens that are engaged and aware
  - Owning up to our Arabic(s)

# Basic Premise

- Goal of the UAE Arabic Project
  - Preservation of the Arabic language
- Why
  - Passion: National pride/identity/character/ culture/religion/connection to heritage, etc.
  - Utility: Language => information => knowledge =>power
    - Namely, Independence from other powers

# But which Arabic(s)?

- The dialects are as good as any other form of Arabic (MSA)
  - We can come up with a standard writing convention for any dialect, no biggie (Maltese)
  - They satisfy the 2 possible whys
    - *Passion: National pride/identity/character/culture/religion/ Connection to heritage*
    - *Utility: Language => information => knowledge =>power*
      - *Namely, Independence from other powers*

# But which Arabic(s)?

- The dialects are as good as any other form of Arabic (MSA)
  - We can come up with a standard writing convention for any dialect, no biggie
  - They satisfy the 2 possible whys
    - *Passion: National pride/identity/character/culture/religion/ Connection to heritage*
    - *Utility: Language => information => knowledge =>power*
      - *Namely, Independence from other powers*
- BUT
  - Power is in *numbers* and *quality* of people who speak (can deal/handle) that language (maybe easier to have a unifying currency )
    - Hence MSA is the preferred variety for the UAE project

# Accordingly MSA

- Divide the problem into smaller ones and attack
  - Preserve a snapshot of the language (corpora)
    - Understand that snapshot really well
  - Increase the number and quality of people who actually use that variety via education/media/ economic utility
  - Transform Arabic into a powerful language with high utility
    - There are inter-dependencies in this picture

# Founding Step

- Preserve a large balanced corpus of MSA as it stands at several intervals a la the BNC and the ANC (*ideological considerations notwithstanding*)
- Annotate it with all levels of relevant linguistic information
- Create an associated computational dictionary a la COBUILD for example
- Leverage existing efforts at various locations: e.g. Princeton, LDC and ELRA

# Increase number & quality of speakers of MSA

- Current impediments
  - Education
    - Issues with curricula, methodologies, etc.
    - Low readership
  - Media
    - MSA is not cool enough (distance from Mexican)
    - Low media content
  - Economic Utility
    - Mostly in English
- One vision: technology can help (specifically CL/NLP)

# Goals of Computational Linguistics

- Model Human Language Processing
- Analyze Human Language
- Facilitate Human Language Communication via Automated Tools

# Natural Language Processing (NLP)

- NLP (*aka* computational linguistics) is the field of automatic processing of natural spoken and written language
- NLP deals mainly with unstructured data
- NLP is at the intersection of several fields
  - computer science, linguistics, machine learning and artificial intelligence, statistics, among others.
- NLP application technologies include
  - machine translation (MT)
  - information retrieval (IR)
  - Information extraction (IE)
  - automatic summarization (AS)
  - automatic speech recognition (ASR)
  - optical character recognition (OCR)
- The research skills acquired are directly transferrable to other sub fields within computer science such as Bio-informatics
- NLP is defined by quantitative evaluation metrics (crucial)

# Objectives

- Take people's everyday language (the way they speak/write) and do useful things with it
- Useful things such as:
  - Translate from one language to another
  - Extract relevant information for a task (distillation, summarization, track opinions, gage people's sentiments towards something/someone)
  - Information retrieval (google/yahoo/bing)
  - Improve pedagogical systems
  - Etc….

# The Challenge

- Language is complex with infinite possible constructions
- Good news is that there are patterns as the symbol set is finite, but the patterns are latent
- Availability of raw data but not explicitly annotated/marked

# Columbia Arabic Dialect Modeling Group (CADIM)

- Founded February 2005
- Natural Language Processing with a focus on Arabic language technologies
- Academic Personnel include Research Scientists, Postdocs and Students
- CADIM members have over 170 international publications combined and numerous software releases, over $5m in grant funding from DARPA, NSF, IARPA, and private industry

# Why CADIM focus on Arabic Language Technology?

- Most significant NLP research takes place in the USA, Europe and Japan
- NLP focus is predominantly on English
- The amount and type of research in Arabic NLP is still limited in scope, focus, and practitioners
- Arabic poses challenges to NLP in general where the lessons learned could be transferred to other languages (Urdu, Chinese, etc), pushing the edge of science

# CADIM Research Thrusts

- Automatic Machine Translation
  - From and to Arabic with all its dialectal variants
- Text Analytics
  - Multilingual and cross lingual information extraction, data mining from unstructured data, summarization and distillation in different media
- Language Pedagogy
  - Aiding educators and classroom language teaching in Arabic and other languages
- Dialogue Systems
  - Dialectal dialogue applications

# Approaching the challenge

- Divide & Conquer
  - Break the problem into smaller problems
- Throw state of the art techniques at the smaller problems

# Where are we today

- Basic Tools for processing MSA (newswire) are in the high 90's
- Machine translation needs a lot of work, but getting there

# How can technology help MSA Education?

- Create better tools that facilitate the learning of MSA and CLA in the classroom
- Create better printed material (e.g. consistent, no typographical errors)
- Aggregate resources for educational purposes: e.g. translation of scientific writing and fiction into Arabic and making it readily available; creating large (online) corpora; creating large machine readable dictionaries, etc.

# Educational Tools

كتير بشوف مواضيع شو هدفك
بالحيا وشو حلمك وشو أمنيتك
والأ ة بتكون بدي اتخرج واشتغل
اجستير ودكتوراه بدي
ب ولاد وعيش بجو أسري
كله مودة بدي كون حالي بدي اعمل
مشروع خاص فيني بدي سافر وهاجر
واشتغل مارح اسألكن هالأسئلة بس
بدي اسأل سؤال غير...لوين بدك
توصل؟وتوصلي؟ايمت رح تقول خلص
وصلت ايمت رح تحس حالك مرتاح
وماتطالب حالك انو توصل اكتر وين
بدك توصل وهل انت رح تكون سعيد؟؟

Click-to-learn capabilities

| Plural | Fem. Sing. | Masc. Sing. | |
|---|---|---|---|
| Past Tense | | | |
| شفنا | شفت | شفت | 1st |
| شفتو | شفتي | شفت | 2nd |
| شافو | شافت | شاف | 3rd |
| Present Tense | | | |
| بنشوف | بشوف | بشوف | 1st |
| بتشوفو | بتشوفي | بتشوف | 2nd |
| بيشوفو | بتشوف | بيشوف | 3rd |
| Future Tense | | | |
| حنشوف | حشوف | حشوف | 1st |
| حتشوفو | حتشوفي | حتشوف | 2nd |
| حيشوفو | حتشوف | حيشوف | 3rd |

# Educational Tools

كتير بشوف مواضيع شو هدفك
بالح ة وشو حلمك وشو أمنيتك
و وبة بتكون بدي اتخرج واشتغل
ماجستير ودكتوراه بدي
يب ولاد وعيش بجو أسري
حلة مودة بدي كون حالي بدي اعمل
مشروع خاص فيني بدي سافر
وهاجر واشتغل مارح اسألكن
هالأسئلة بس بدي اسأل سؤال
غير...لوين بدك توصل؟وتوصلي؟ايمت
رح تقول خلص وصلت ايمت رح
تحس حالك مرتاح وماتطالب حالك
انو توصل اكتر وين بدك توصل وهل
انت رح تكون سعيد؟؟

Click-to-learn capabilities

| Dialect ID | Levantine (70%) Egyptian (20%) MSA (10%) |
|---|---|
| Lemma | شاف |
| MSA Equivalent | رأى نظر لاحظ |
| English | See, observe, notice |

# Automatic Diacritization

- حرصاً منا على تقديم الخدمة الأفضل للبث المباشر بموقع الجزيرة نت، فقد انطلقت خدمة تجريبية جديدة للبث المباشر تتيح لجمهورنا الكريم مشاهدة قناة الجزيرة.

- حرْصَا منّا عَلَىَ تَقْديمْ الخْدْمَة الأفْضَلِ لَلَبثَ المُبَاشرُ بمَوْقعِ الجْزيرَة نت، فَقَدْ اَنْطَلَقْتُ خدْمَة تجْرِيْبِيَّةَ جَدِيْدَةً لَلَبِثَ المُبَاشِرُ تُتِيَحُ لجُمْهُورَنَا الْكَرِيمْ مُشَاهَدَةِ قَنَاةِ الجْزِيرَةِ.

# Address the low readability issue

- Investigate diacritization scientifically
  - Current scientific research suggests that full diacritization is an impediment
  - No diacritization is quite bad as well
  - Possible solutions (partial diacritization)
    - Scientific research between psycholinguists, neurolinguists, computational linguists and L2 educational structures
- Based on results, change printing strategies and editorial guidelines

# How can technology help Social Media?

- Increase Arabic content on the web via translation
- Encourage bloggers and writers to write in Arabic online
- Exploit social media for economic benefit (tie social media to ROI): tracking people's opinions/sentiments on products/events/people; recommender systems; multimodal systems (video, speech, text); etc.
- Building better engaging civic structures through e-government
- Crowd sourcing for policy making such as polling on relevant social and political issues early on in the political process

# How can technology help with economic utility?

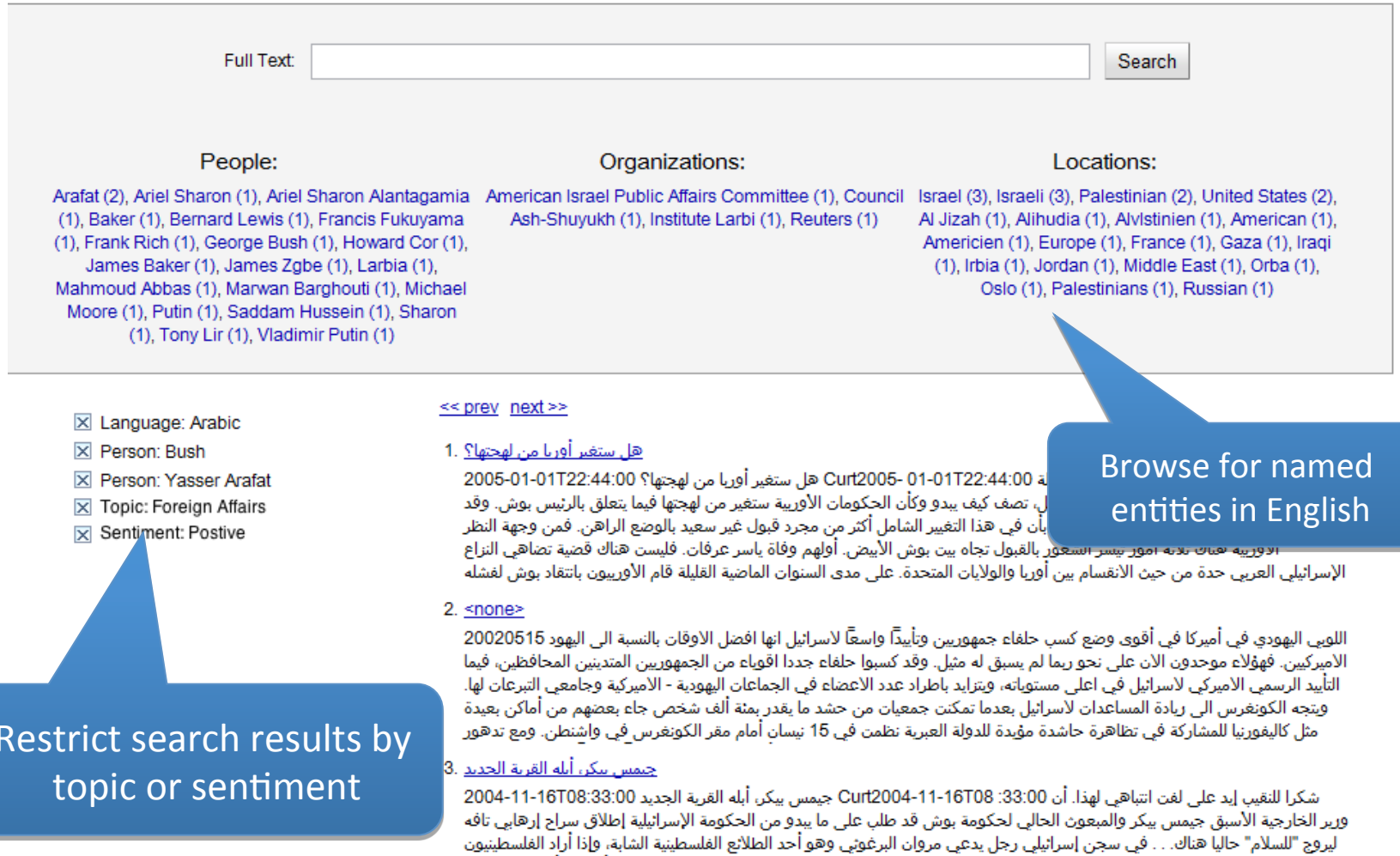
- Create the our Google, Yahoo, IBM
- Spin-offs and incubations

- Make Arabic cool! Tie it to jobs and utility

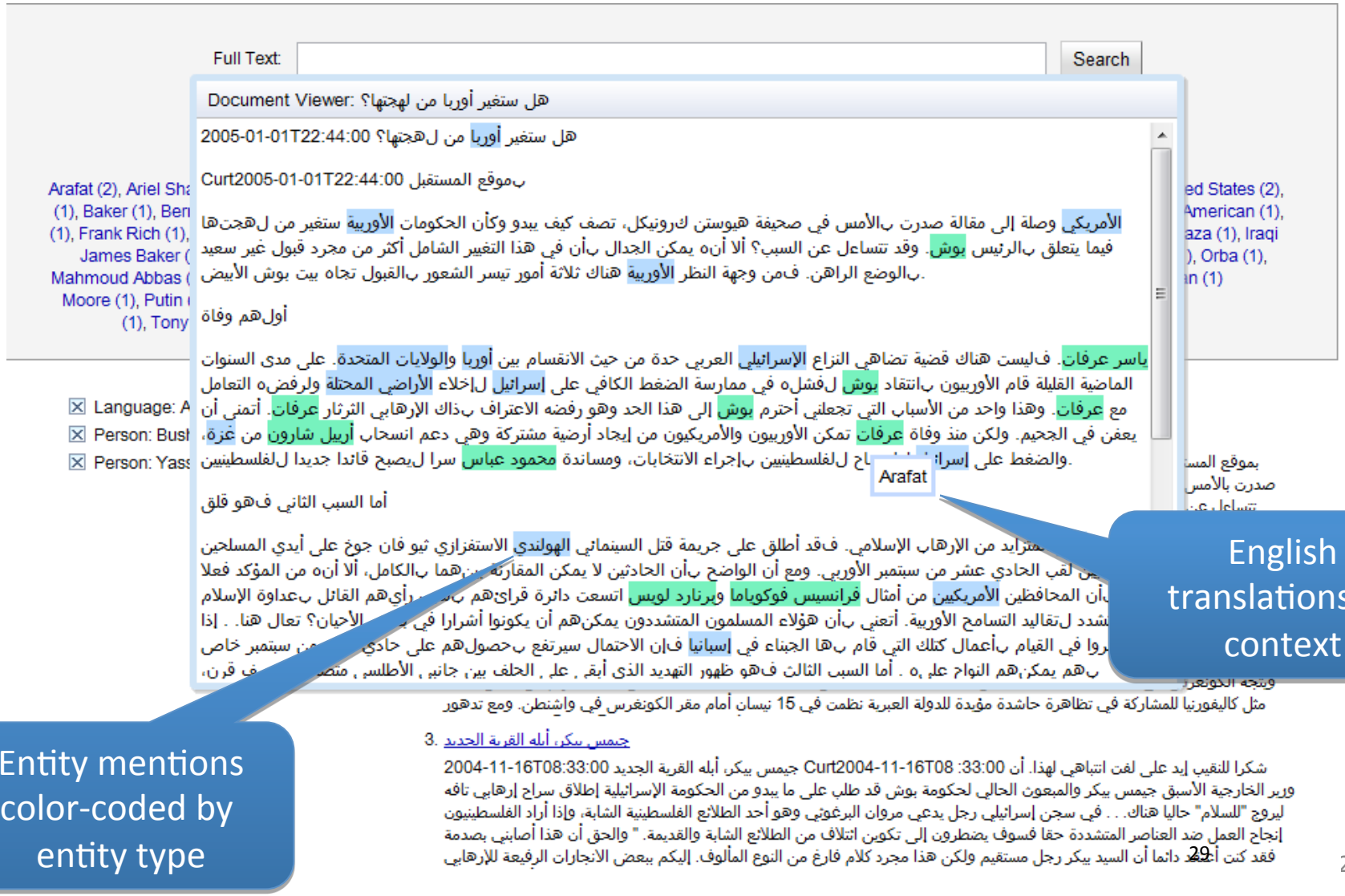
# Transform Language into Power

- Imagine you are a decision maker
- You press on a button
- You get a distilled report of your competitors position in the market or your opponents reputation, or …etc. based on articles, memos, blogs, videos, opinions
  - Yes this is doable (but not in Arabic yet)

# Multi-Lingual Faceted Search



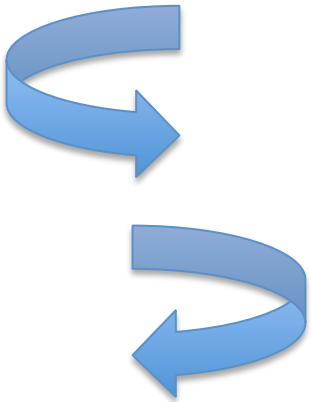
# Highlight Information of Interest



# What is missing for the Arabic language to become information/power?

- We will have to seriously address the ubiquity of the dialects issue (it is a reality)
- More scientists engaging the Arabic language processing problems

# Chicken & Eggs



*Not going to discuss here*

# Reality: Arabic Sample Data

| Data Source | Example |
|---|---|
| Newswire<br>MSA only | واكد لليوم الثانى ان "الجهود مستمره الى الامام" من اجل مواصلته الحوار الوطنى بخصوص عملية السلام.<br>*And he emphasized for the second day that "efforts are continuing forward" to resume the national dialogue on the peace process.* |
| Broadcast<br>MSA+some DIA | علشان كده هي بتتفاعل مع ما يحدث وتجد إلزاما عليها أن تنبه الشعب العربي إلى حقيقة ما يدور بالمفاوضات<br>*'cause o' this it's interactin' with what is happening and it finds it necessary to awaken the Arab people to the truth of what is happening in the negotiations* |
| CTS, news groups & blogs<br>more DIA | بالعاكس عادي بس لأني متأكد إني بعرفكيش عشان هيك بحكي لك إنتي مخربطة<br>*no problem, but since I am sure I don't know you, that's why I am telling you you're confused.* |

# Reality: Using MSA Tools

| | **MSA data** | **Broadcast data** | **Dialect data** |
|---|---|---|---|
| **TOK** | *99.6* | *98.1* | *97* |
| **POS** | *97.3* | *95* | *72* |
| **BPC** | *93* | *89* | *82* |

# Reality: Tools made for MSA fail on Arabic dialects

| Arabic Variant | Arabic Source Text | Google Translate |
|---|---|---|
| Egyptian | الكهربا اتقطعت، ليه كده بس؟ | Atqtat electrical wires, Why are Posted? |
| Levantine | شكلو مفيش كهربا ، ليش هيك؟ | Cklo Mafeesh ابرهك, Lech heck? |
| Iraqi | شو ماكو كهرباء، خير؟ | Xu MACON electricity, good? |
| MSA | لايوجد كهرباء، ماذا حصل؟ | Does not have electricity, what happened? |

# Reality: Dialect Switching

MSA

LEV

MSA and Dialect mixing in speech

- phonology, morphology and syntax

لا أنا ما بعتقد لأنه عملية اللي عم بيعارضوا اليوم تمديد للرئيس لحود هم اللي طالبوا بالتمديد للرئيس الهراوي وبالتالي موضوع منه موضوع مبدئي على الأرض أنا بحترم أنه يكون في نظرة ديمقراطية للأمور وأنه يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية وبعتقد إنه الكل في لبنان أو أكثرية ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع إنجازات العهد يعني نعم نحكي عن إنجازات العهد لكن هل النظام في لبنان نظام رئاسي النظام في لبنان من بعد الطائف ليس نظام رئاسي وبالتالي السلطة هي عمليا بيد الحكومة مجتمعة والرئيس لحود أثبت خلال ممارسته الأخيرة بأنه لما بيكون في شخص مسؤول في منصب معين وأنا عشت هذا الموضوع شخصيا بممارستي في موضوع الاتصالات لما بياخد مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو إلى جانبه إنما مش مطلوب من رئيس جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه بقى في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صح عليه تثمير جهود الوطنية الشاملة كي يظل في مصالحة وطنية كي يظل في توافق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما يترك المسار يروح باتجاه الخطأ نعم إنما خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي مشيوا معه وآمنوا فيها التزموا فيها أنا أثبت خلال الأربع سنوات بالممارسة الحكومية أني التزمت فيها ولما التزمنا بهذا الموضوع كان الرئيس لحود إلى جنبنا في هذا الموضوع، أما الموضوع الديمقراطي أنا بتفهم تماما هذا هالوجهة النظر بس ما ممكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب ديمقراطي ضمن المجلس والتصويت إلى ما هنالك لرئيس جمهورية بولاية ثانية هو مسح هيئة في جوهر الديمقراطية هذا بالأقل يعني قناعتي في هذا الموضوع.



Aljazeera Transcript http://www.aljazeera.net/programs/op_direction/articles/2004/7/7-23-1.htm

# Reality: Dialect Switching

MSA and Dialect mixing in speech

- phonology, morphology and syntax

**MSA-LIKE LEV**

**MSA**

**LEV**

لا أنا ما بعتقد لأنه عملية اللي عم بيعارضوا اليوم تمديد للرئيس لحود هم اللي طالبوا بالتمديد للرئيس الهراوي وبالتالي موضوع منه موضوع مبدئي على الأرض أنا بحترم أنه يكون في نظرة ديمقراطية للأمور وأنه يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية وبعتقد إنه الكل في لبنان أو أكثرية ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع إنجازات العهد يعني نعم نحكي عن إنجازات العهد لكن هل النظام في لبنان نظام رئاسي النظام في لبنان من بعد الطائف ليس نظام رئاسي وبالتالي السلطة هي عمليا بيد الحكومة مجتمعة والرئيس لحود أثبت خلال ممارسته الأخيرة بأنه لما بيكون في شخص مسؤول في منصب معين وأنا عشت هذا الموضوع شخصيا بممارستي في موضوع الاتصالات لما بياخد مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو إلى جانبه إنما مش مطلوب من رئيس جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه بقى في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صح عليه تثمير جهود الوطنية الشاملة كي يظل في مصالحة وطنية كي يظل في توافق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما يترك المسار يروح باتجاه الخطأ نعم إنما خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي مشيوا معه وآمنوا فيها التزموا فيها أنا أثبت خلال الأربع سنوات بالممارسة الحكومية أني التزمت فيها ولما التزمنا بهذا الموضوع كان الرئيس لحود إلى جنبنا في هذا الموضوع، أما الموضوع الديمقراطي أنا بتفهم تماما هذا هالوجهة النظر بس ما ممكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب ديمقراطي ضمن المجلس والتصويت إلى ما هنالك لرئيس جمهورية بولاية ثانية هو مسح هيئة في جوهر الديمقراطية هذا بالأقل يعني قناعتي في هذا الموضوع.



Aljazeera Transcript http://www.aljazeera.net/programs/op_direction/articles/2004/7/7-23-1.htm

# Reality: Dialectal Impact on MSA

- Code switching in written MSA
- Dialectal lexical and structural uses
  - Example Newswire Alnahar newspaper (ATB3 v.2)

فأخذ على خاطر الأخوان ومن حقهم ان يزعلوا

*f>x* ElY xATr* AlAxwAn wmn hqhm An *yzElw*

*then_was-taken upon self the-brothers and-from right-their to be-angry*

'they were upset, and they had the right to be angry'

# Possible Short Term Projects (3-5 yrs) Doable today

- Create a balanced annotated corpus of MSA
- Create a large comprehensive contemporary MSA dictionary that covers all the unique types in the corpus
- Create a large dialect to MSA dictionary
- Create basic robust MSA tools for different genres of text
- Create tools to automatically fix typographical errors and (partially) diacritize text

# Possible Medium Term Projects

- Addressing the readability issue
- Addressing the code switching phenomena
- Machine translation of very different genres and domains robustly
- Handling the dialectal phenomena both linguistically and computationally (expand on our constructed corpus)

# Possible Components/needs: Parallel tracks (?)

- Scientific research in Arabic NLP
- Scientific research in Arabic linguistics
- Engagement with educational community
- Engagement with economic investors
- Engagement with political entities
- Money, money, money!

A vision with synergy as an underlying premise
A community of people with interest
A/several VISIONARY/IES with FUNDING

# Results if carried out

- Prepare world-class researchers through both formal and informal research education
- Create a ***sustainable*** educational and research environment
- Advance state-of-the-art Arabic language technologies
- Build strong research community with skills for innovation and technological creativity in the Middle East and North Africa (MENA) region
- The skill set is directly transferable to other domains (bio technology)
- Transfer technology via spin-offs, licensing

# Final word

- This was a roadmap for real science that could be our own
- A paradigm shift!

Let the journey begin?