Parsing Arabic Dialects

Owen Rambow
Columbia University
rambow@cs.columbia.edu

Team

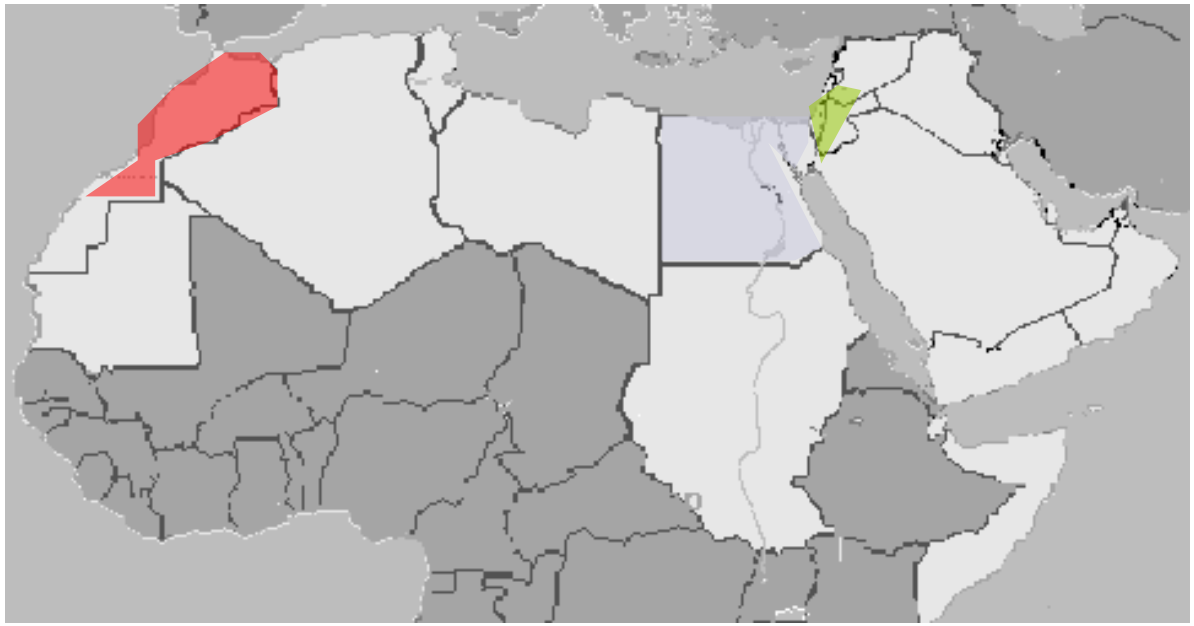
- Senior Members
  - David Chiang — U of Maryland
  - Mona Diab — Columbia
  - Nizar Habash — Columbia
  - Rebecca Hwa — U of Pittsburgh
  - Owen Rambow — Columbia
  - Khalil Sima'an — U of Amsterdam
- Grad Students
  - Roger Levy — Stanford
  - Carol Nichols — Pittsburgh

Team (ctd)

- Undergrads
  - Vincent Lacey — Georgia Tech
  - Safiullah Shareef — Johns Hopkins
- Externals
  - Srinivas Bangalore, — AT&T Labs -- Research
  - Martin Jansche — Columbia
  - Stuart Shieber — Harvard
  - Richard Sproat — U of Illinois at UC
  - Bill Young — CASL/U of Maryland

Overview

- Team
- **Problem: Why Parse Arabic Dialects?**
- Proposed Approaches
- Accomplishments to Date
- Plan for Workshop
- Summary



lam jaʃtari nizār ţawilatan ʕadīdatan لم يشتر نزار طاولة جديدة

didn't buy Nizar table new

nizār maʃtarāʃ ţarabēza gidīda نزار ماشتراش طربيزةجديدة

nizār maʃtarāʃ ţawile ʕdīde نزار ماشتراش طاولة جديدة

nizar maʃrāʃ mida ʕdīda نزار ماشراش ميدة جديدة

Nizar not-bought-not table new

Factors Affecting Dialect Usage

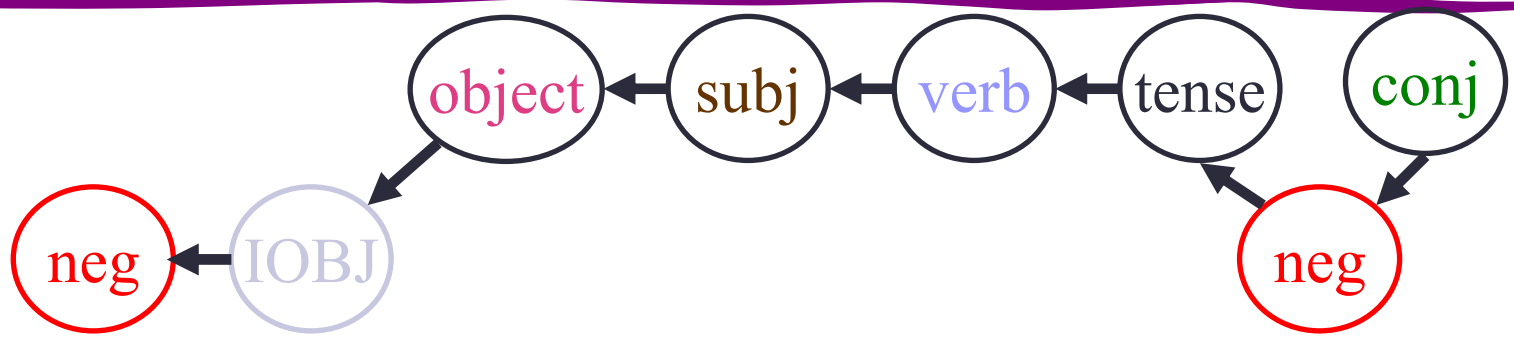
- Geography (continuum)
- City vs village
- Bedouin vs sedentary
- Religion, gender, …

$\Rightarrow$ Multidimensional continuum of dialects

Lexical Variation

English	table	cat	of	(I) want	there is	there isn't
MSA	Tāwila	qiTTa	*idafa*	ʾuridu	yūjadu	lā yujadu
Moroccan	mida	qeTTa	dyāl	bġit	kāyn	mā kāynš
Egyptian	Tarabēza	ʾoTTa	bitā3	3āwez	fī	mafīš
Syrian	Tāwle	bisse	taba3	biddi	fī	mā fi
Iraqi	mēz	bazzūna	māl	ʾarid	aku	māku

Morphological Variation
Verb Morphology



MSA

ولم تكتبوها له

walam taktubūhā lahu

wa+lam taktubū+hā la+hu

and+not_past write_you+it for+him

EGY

وماكتبتوهالوش

wimakatabtuhalūʃ

wi+ma+katab+tu+ha+lū+ʃ

and+not+wrote+you+it+for_him+not

And you didn't write it for him

Dialect Syntax: Word Order

- Verb Subject Object
  كتب الاولاد الاشعار
  wrote.masc the-boys the-poems (MSA)
- Subject Verb Object
  الاولاد كتبو الاشعار
  the-boys wrote.masc.pl$_{I}$ the-poems (LEV, EGY)

	Most common word order	Full agreement in VSO	Full agreement in SVO
MSA	VSO	no	√
Dialects	SVO	√	√

Dialect Syntax: Noun Phrases

- Possessives
  - Idafa construction
    - **Noun1 Noun2**
    - ملك الاردن
      king Jordan
      the king of Jordan / Jordan's king
  - Dialects have an additional common construct
    - **Noun1 *<particle>* Noun2**
    - LEV: الملك تبع الاردن the-king *belonging-to* Jordan
    - <particle> differs widely among dialects
- Pre/post-modifying demonstrative article
  - MSA: هذا الرجل this the-man *this man*
  - EGY: الراجل ده the-man this *this man*

Code Switching: Al-Jazeera Talk Show

MSA and Dialect mixing in formal spoken situations

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 بترك 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

Why Study Arabic Dialects?

- There are **no native speakers of MSA**
- **Almost no** native speakers of Arabic are able to sustain continuous spontaneous production of spoken MSA
- This affects **all** spoken genres which are not fully scripted: conversational telephone, talk shows, interviews, etc.
- Dialects also in use in new media (newsgroups, blogs, etc)

Arabic Dialects: Computational Resources (1)

- Transcribed speech/transcript corpora
  - **Levantine** (LDC), Egyptian (LDC), Iraqi, Gulf, …
- Very little other unannotated text
  - Online: Blogs, newsgroups
  - Paper: Novels, plays,soap opera scripts, …
- Teensy treebanks
  - **Levantine**, LDC for this workshop with no funding
  - JUST FOR EVALUATION

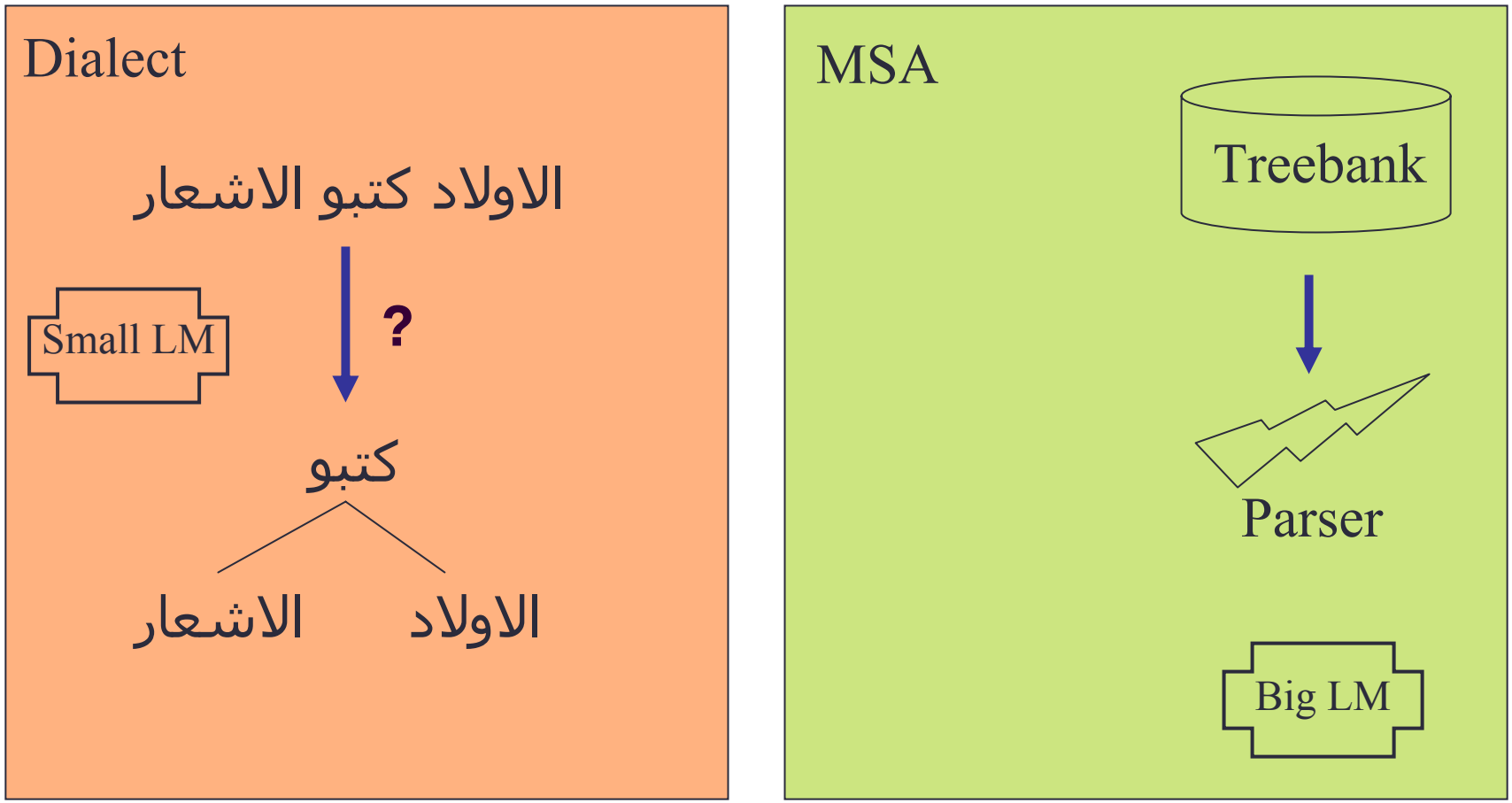
Arabic Dialects: Computational Resources (2)

- Lexicons
  - CallHome Egyptian Arabic monolingual lexicon (LDC)
  - Egyptian-MSA and **Levantine-MSA dictionary** (Columbia, NSF funding)
  - Wordlists (CASL, web, published sources)
- Morphological resources
  - CallHome Egyptian Verb transducer (LDC)
  - Columbia University Arabic Dialect Project: MAGEAD: Pan-Arab Morphology, only MSA so far (ACL workshop 2005)
  - Buckwalter **morphological analyzer for Levantine** (LDC, under development)
- Contrast to MSA: huge corpora, treebank, lexicons, morphological analyzers, taggers, chunkers, parsers, MT system, ASR systems, …

Overview

- Team
- Problem: Why Parse Arabic Dialects?
- **Proposed Approaches**
- Accomplishments to Date
- Plan for Workshop
- Summary

Parsing Arabic Dialects: Problem



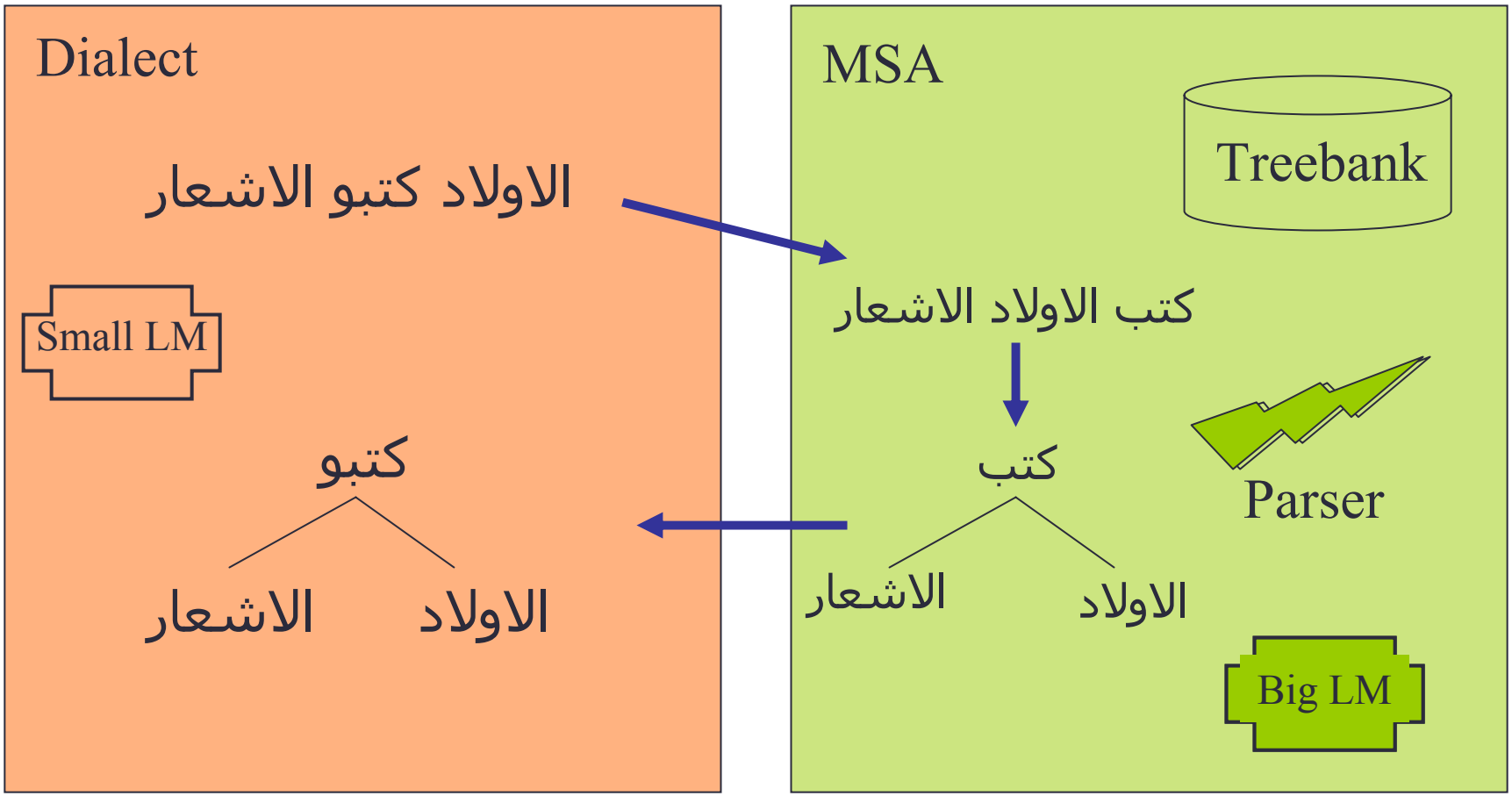
Parsing Arabic Dialects

- Many different dialects
- Dialects are spoken, few written resources
- Code switching
- Conclusion: Can’t assume we will get treebanks or even large unnatoated corpora for each dialect
- What to do?

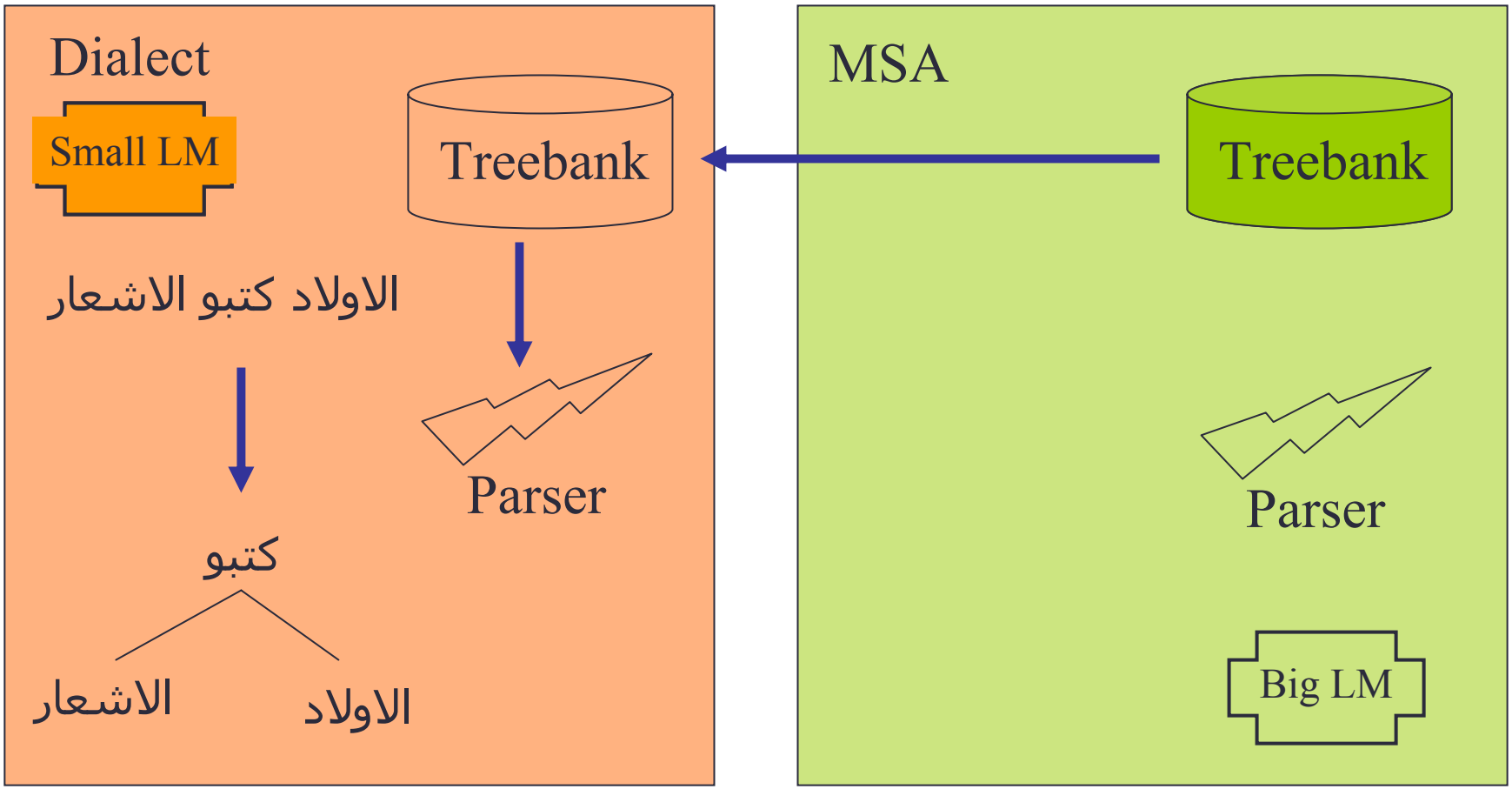
Parsing Arabic Dialects

- Idea: use rich resources for MSA, adapt for dialects
- Exploit:
  - Similarity MSA-dialects
  - Explicit knowledge we have about differences
- Three approaches

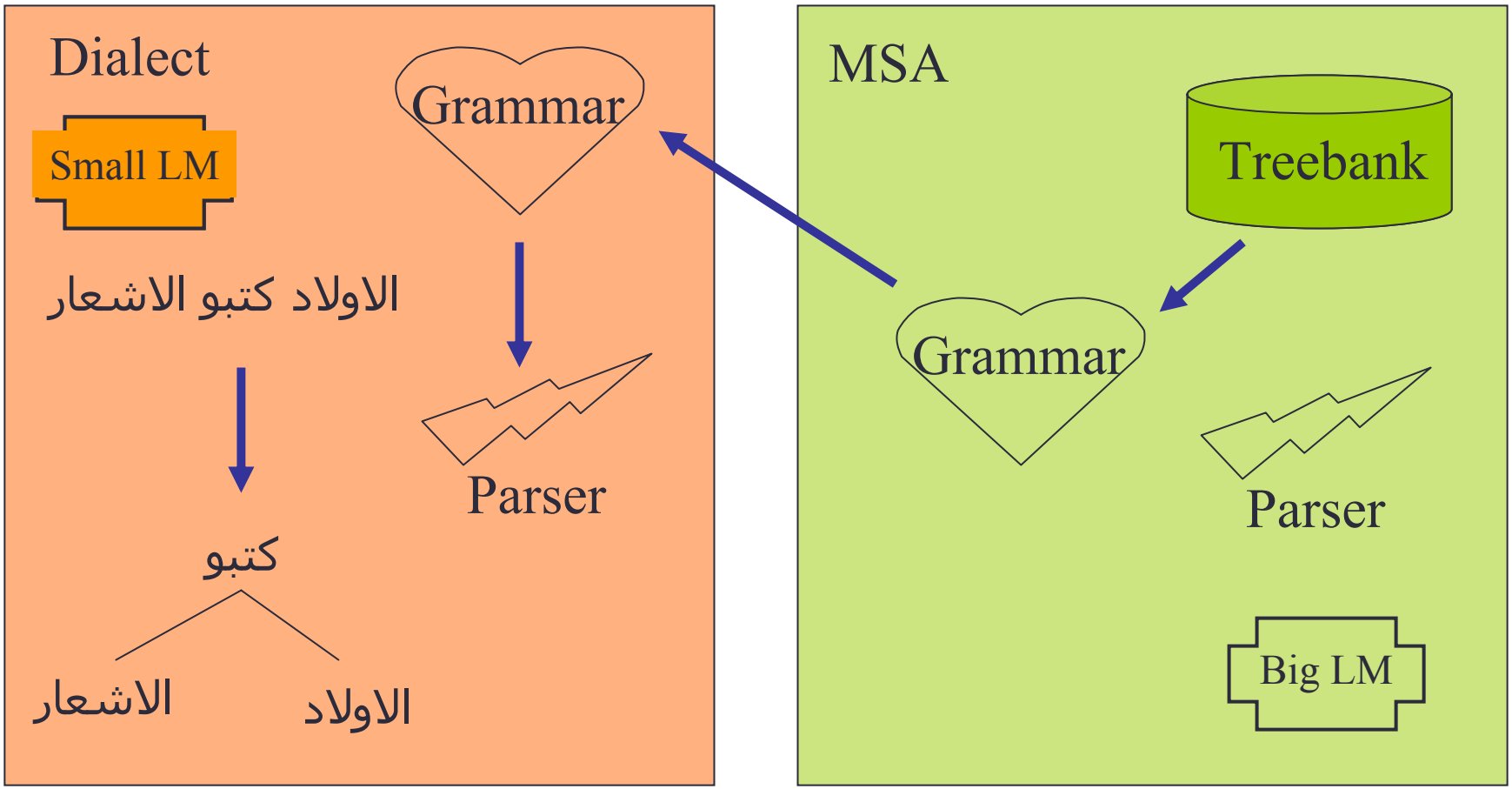
Proposed Solution 1: Dialect Sentence Transduction



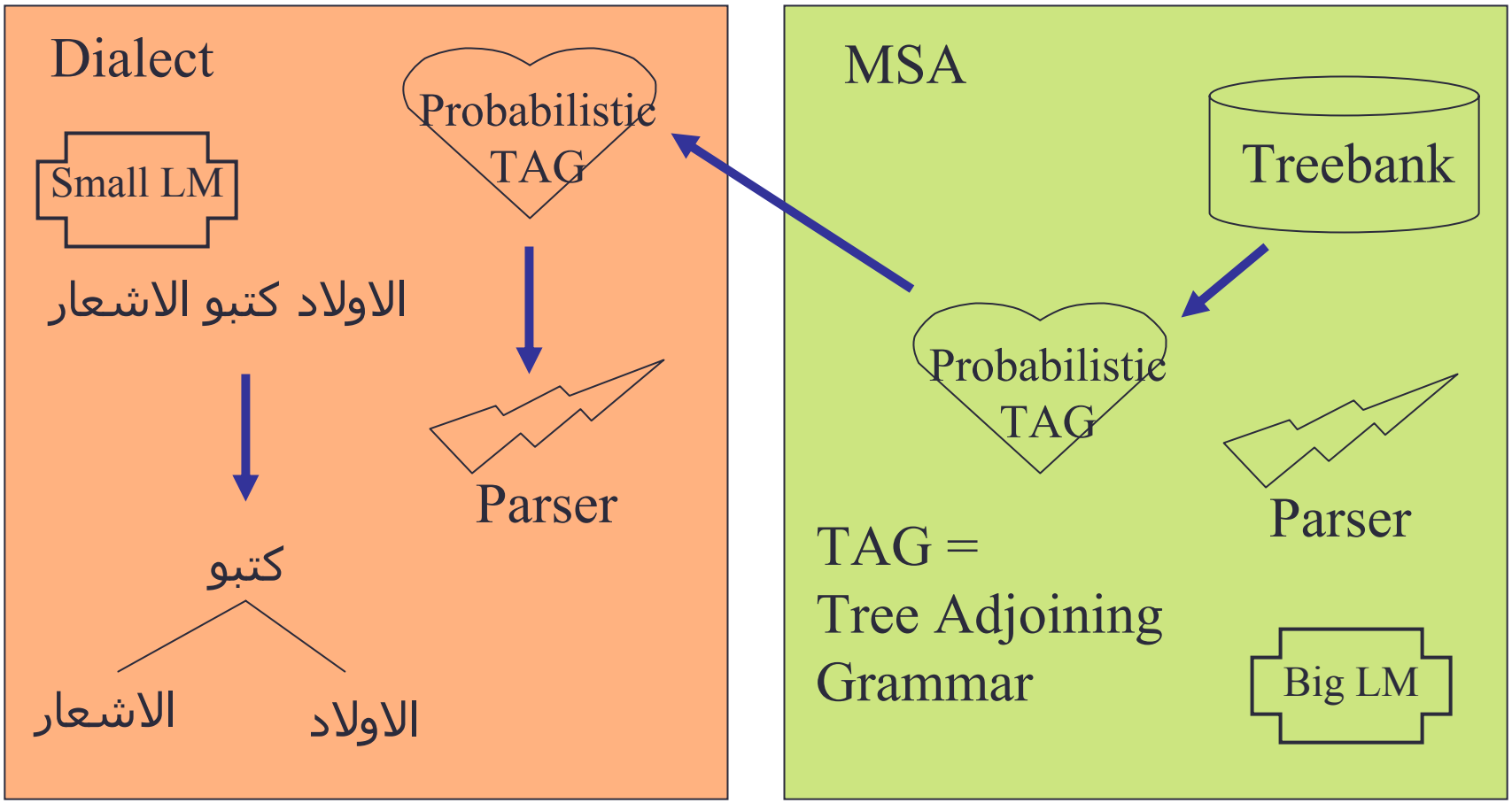
Proposed Solution 2: MSA Treebank Transduction



Proposed Solution 3: MSA Grammar Transduction



Proposed Solution 3: MSA Grammar Transduction



Overview

- Team
- Problem: Why Parse Arabic Dialects?
- Proposed Approaches
- **Accomplishments to Date**
- Plan for Workshop
- Summary

Accomplishments so Far: All Three Approaches

- Baselines: using MSA parsers on Levantine (done/ongoing)
- Seed MSA/Levantine dictionary (by hand, ongoing)
- Finding Levantine text on the web (ongoing)
- Estimating language models for MSA and Levantine (done/ongoing)
- Estimating lexical translation probabilities without a parallel corpus (detailed proposal)

Accomplishments so Far: Dialect-to-MSA Sentence Transduction

- Sentence transducer (implemented)
- Adapt sentence transducer to Levantine (ongoing)
- Backmapper for trees MSA-to-dialect (detailed proposal)

Accomplishments so Far: MSA-to-Dialect Treebank/Grammar Transduction

- Grammar extracted from MSA treebank (more work needed)
- Tree transformation rule formalism and rules for MSA-to-Levantine (implemented, under revision)
- Synchronous grammar (with explicit or hidden grammar): parallel parsing without a parallel corpus (detailed proposal)
- Treebank transfer: estimating parameters without a parallel corpus (detailed proposal)

Work Plan

- Implement all three proposals by end of week 2
- Determine how to proceed

Summary

- Continuum of dialects
- People communicate spontaneously in Arabic dialects, not in MSA
- So far no computational work on dialects, almost no resources (not even much unannotated text)
- Do not want ad-hoc solution for each dialect
- Exploit knowledge of differences MSA/dialects