Parsing Arabic Dialects

JHU Summer Workshop

Final Presentation

August 17, 2005

Global Overview

- **Introduction (Owen Rambow)**
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion

Team

- Senior Members

David Chiang	U of Maryland
Mona Diab	Columbia
Nizar Habash	Columbia
Rebecca Hwa	U of Pittsburgh
Owen Rambow	Columbia (team leader)
Khalil Sima'an	U of Amsterdam

- Grad Students

Roger Levy	Stanford
Carol Nichols	U of 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
Otakar Smrz	Charles U, Prague
Richard Sproat	U of Illinois at UC
Bill Young	CASL/U of Maryland

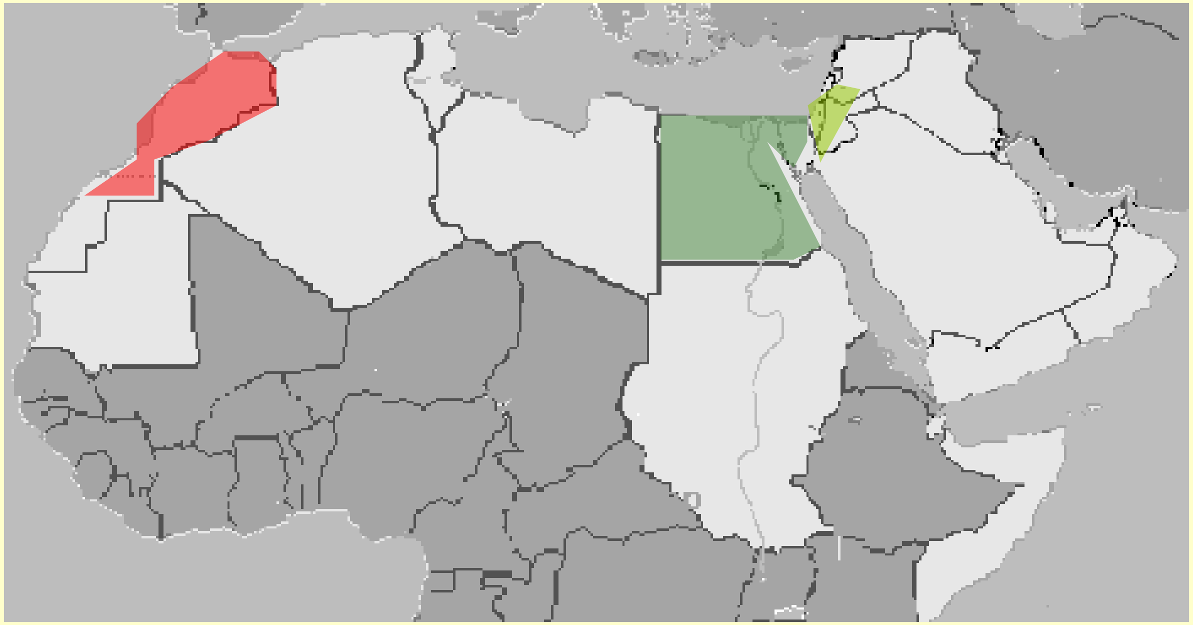
Contact: Owen Rambow, rambow@cs.columbia.edu

Local Overview: Introduction

- Team
- **Problem: Why Parse Arabic Dialects?**
- Methodology
- Data Preparation
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing

The Arabic Language

- Written language: Modern Standard Arabic (MSA)
- MSA also spoken in scripted contexts (news broadcasts, speeches)
- Spoken language: dialects



lam jaʃtari nizār ṭawilatan ζadīdatan — لم يشتر نزار طاولة جديدة

didn’t buy Nizar table new

nizār maʃtarāʃ ṭarabēza gidīda 🟢 نزار ماشتراش طربيزةجديدة

nizār maʃtarāʃ ṭawile ζdīde 🟢 نزار ماشتراش طاولة جديدة

nizar maʃrāʃ mida ζdīda 🔴 نزار ماشراش ميدة جديدة

Nizar not-bought-not table new

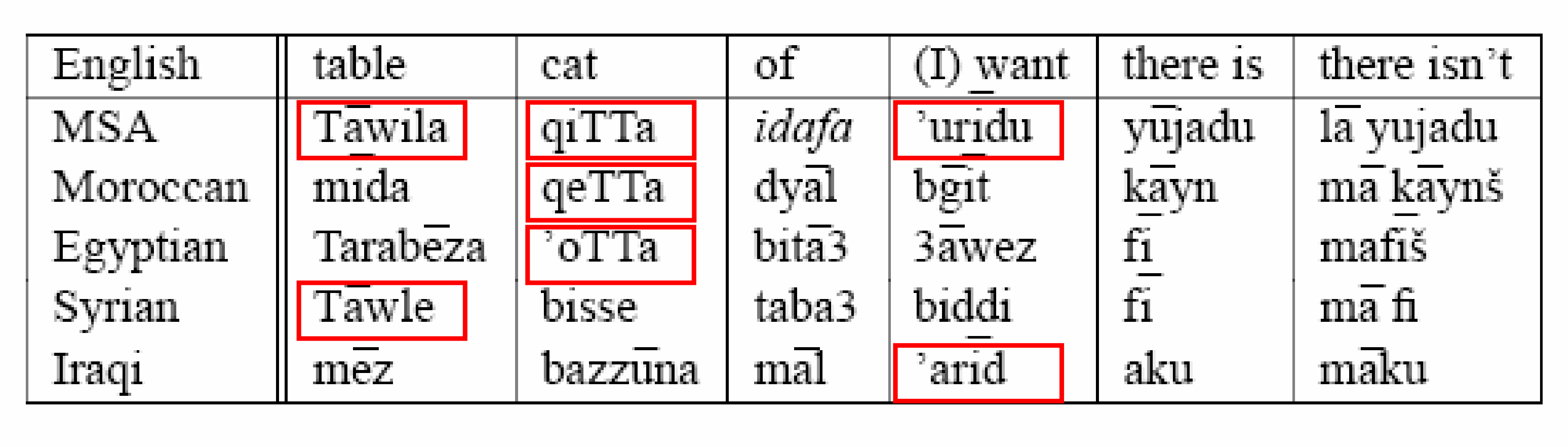
Factors Affecting Dialect Usage

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

⇒ Multidimensional continuum of dialects

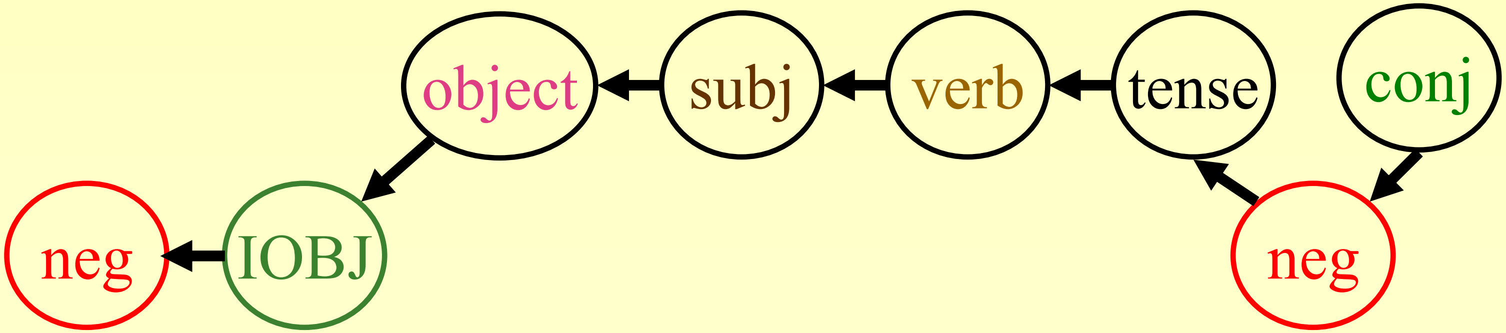
Lexical Variation

- Arabic Dialects vary widely lexically



English	table	cat	of	(I) want	there is	there isn’t
MSA	Tāwila	qiTTa	*idafa*	’urīdu	yūjadu	lā yujadu
Moroccan	mida	qeTTa	dyāl	bgīt	kāyn	mā kāynš
Egyptian	Tarabēza	’oTTa	bitā3	3āwez	fī	mafīš
Syrian	Tāwle	bisse	taba3	biddi	fī	mā fī
Iraqi	mēz	bazzūna	māl	’arīd	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$_l$ the-poems (LEV, EGY)

	VS Order	V	SV Order	Full agreement in VSO	Full agreement in SVO
MSA	35%	30%	35%	no	yes
Dialects	11%	62%	27%	yes	yes

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 بقي 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

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 written media (newsgroups, blogs, etc)
- Arabic NLP components for many applications need to account for dialects!

Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- **Methodology**
- Data Preparation
- Preview of Remainder of Presentation:
    - Lexicon
    - Part-of-Speech Tagging
    - Parsing

Possible Approaches

- Annotate corpora ("Brill Approach")
- Leverage existing MSA resources OUR APPROACH
  - Difference MSA/dialect not enormous: can leverage
  - We have linguistic studies of dialects ("scholar-seeded learning")
  - Too many dialects: even with dialects annotated, still need leveraging for other dialects
  - Code switching: don't want to annotate corpora with code-switching

Goal of this Work

- Goal of this work: show that leveraging MSA resources for dialects is a viable scientific and engineering option
- Specifically: show that using lexical and structural knowledge of dialects can be used for dialect parsing
- Question of cost ($) is an accounting question

Out of Scope

- Tokenization
- Morphological analyzer (but not a morphological disambiguator)
- Speech Effects
  - Repairs and edits
  - Disfluencies
  - Parentheticals
  - Speech sounds

- No standard orthography for dialects
  - Egyptian /mabinjulhalak$/:
    - mAbin&ulhalak$
    - mA bin}ulhAlakS
    - mA binqulhA lak$
    - …
- Issue of ASR interface
- Easy

In Scope

- Deriving bidialectal lexica
- Part-of-speech tagging
- Parsing

Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- **Data Preparation**
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - Parsing

Arabic Dialects: Computational Resources

- Transcribed speech/transcript corpora
  - **Levantine** (LDC), Egyptian (LDC), Iraqi, Gulf, …
- Very little other unannotated text
  - Online: Blogs, newsgroups
  - Paper: Novels, plays,soap opera scripts, …
- Treebanks
  - **Levantine**, LDC for this workshop with no funding
  - INTENDED FOR EVALUATION ONLY
- Morphological resources
  - Columbia University Arabic Dialect Project: MAGEAD: Pan-Arab Morphology, only MSA so far (ACL workshop 2005)
  - Buckwalter **morphological analyzer for Levantine** (LDC, under development, available as black box)

MSA: Computational Resources

- Huge unannoted corpora,
- MSA treebank (LDC)
- Lexicons,
- Morphological analyzers (Buckwalter 2002)
- Taggers (Diab et al 2004)
- Chunkers (Diab et al 2004)
- Parsers (Bikel, Sima’an)
- MT system, ASR systems, …

Data Preparation

- 20,000 words of Levantine (Jordanian) syntactically annotated by LDC
- Removed speech effects, leaving 16,000 words (4,000 sentences)
- Divided into development and test data
- Note: NO TRAINING DATA
- Use morphological analysis of LEV corpus as a standin for true morphological analyzer
- Use MSA treebank from LDC (300,000 words) for training and development
- Contributors: Mona Diab, Nizar Habash

Issues in Test Set

- Annotated Levantine corpus used only for development, testing (no training)
- Corpus developed rapidly at LDC (Maamouri, Bies, Buckwalter), for free (thanks!)
- Issues in corpus:
    - 5% words mis-transcribed
    - Some inconsistent annotations

Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Data Preparation
- **Preview of Remainder of Presentation:**
  - **Lexicon**
  - **Part-of-Speech Tagging**
  - **Parsing**

Bidialectal Lexicons

- Problem:
  - No existing bidialectal lexicons (even on paper)
  - No existing parallel corpora MSA-dialect
- Solution:
  - Use human-written lexicons
  - Use comparable corpora
  - Estimate translation probabilities

Part-of-Speech Tagging

- Problem:
  - No POS-annotated corpus for dialect
- Solution 1: adapt existing MSA resources
  - Minimal linguistic knowledge
  - MSA-dialect lexicon
- Solution 2: find new types of models

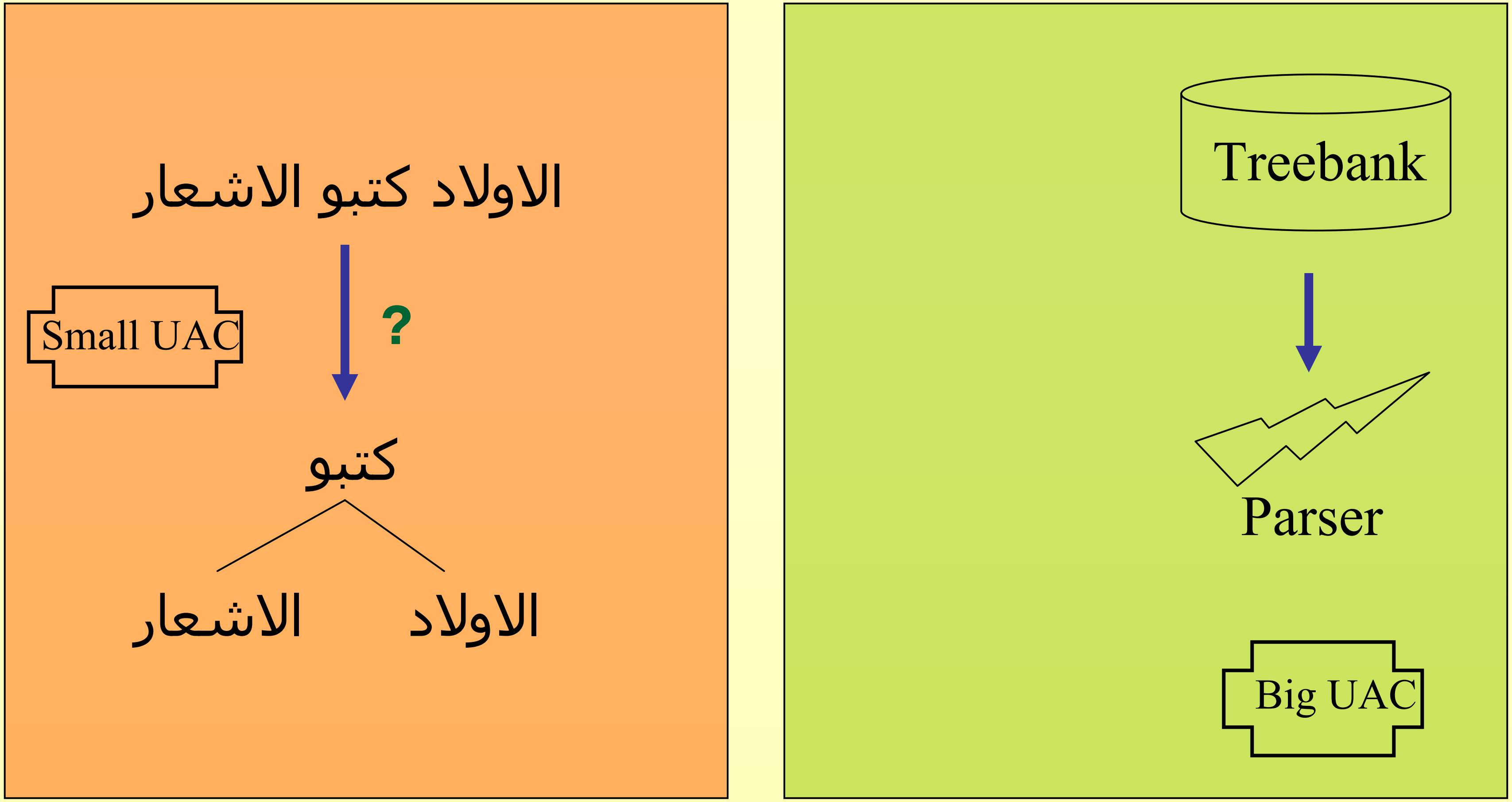
Local Overview: Introduction

- Team
- Problem: Why Parse Arabic Dialects?
- Methodology
- Preview of Remainder of Presentation:
  - Lexicon
  - Part-of-Speech Tagging
  - **Parsing**

Parsing Arabic Dialects:
The Problem

- Dialect -

- MSA -



الاولاد كتبو الاشعار

Small UAC

?

كتبو

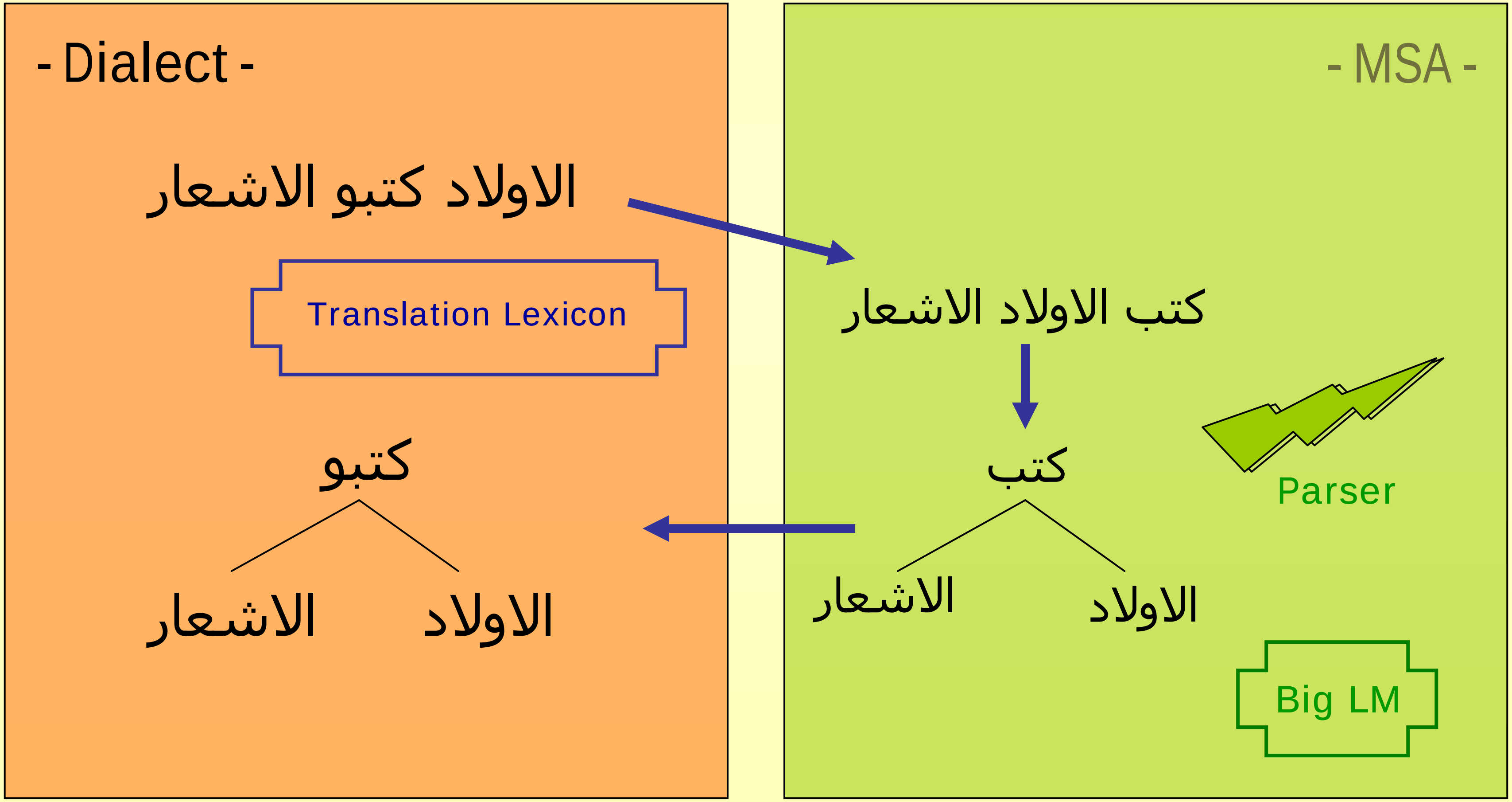
الاشعار الاولاد

Treebank

Parser

Big UAC

Parsing Solution 1:
Dialect Sentence Transduction



- Dialect -

الاولاد كتبو الاشعار

Translation Lexicon

كتبو

الاولاد

الاشعار

- MSA -

كتب الاولاد الاشعار

كتب

الاولاد

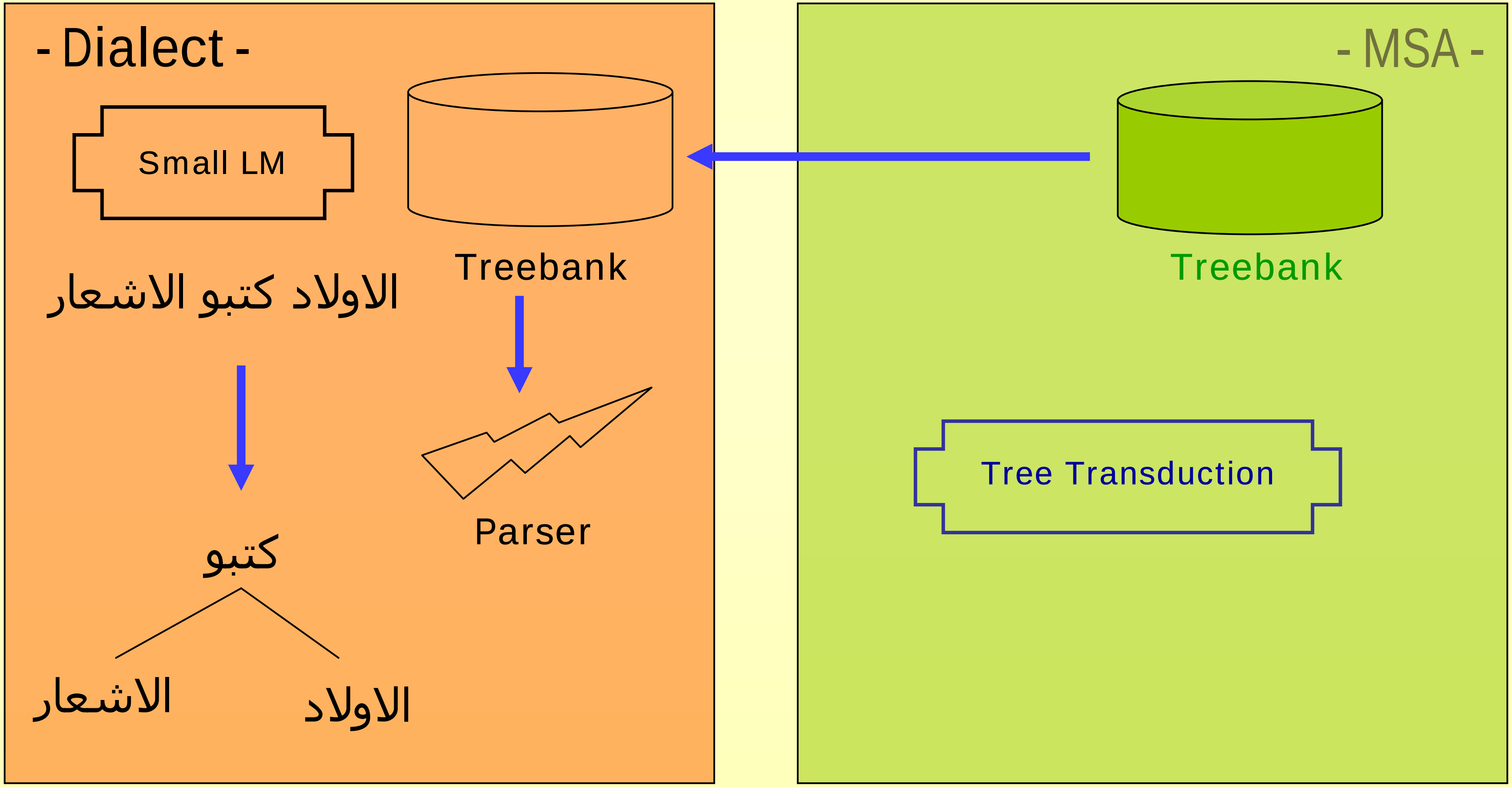
الاشعار

Parser

Big LM

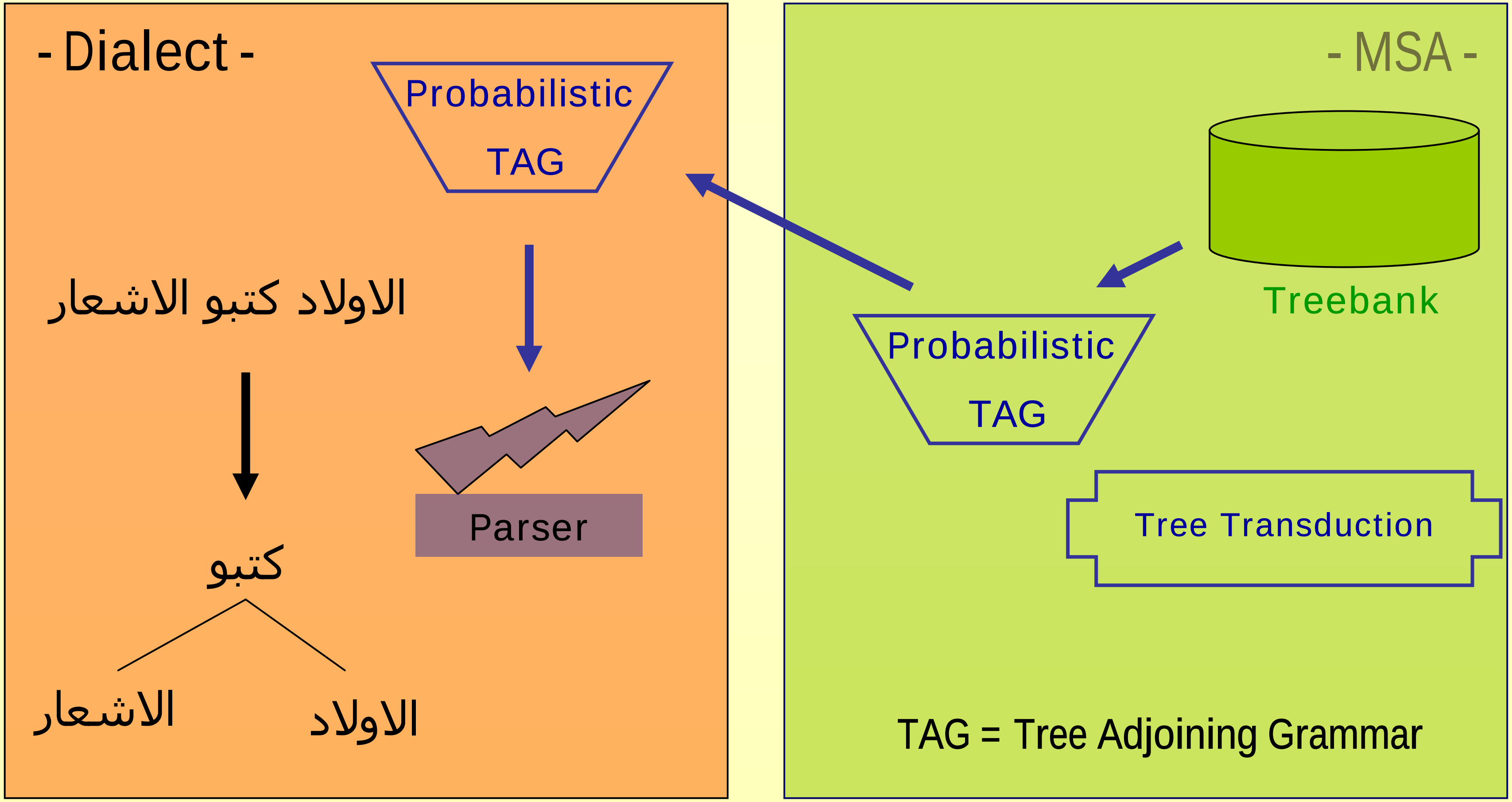
🟦 Workshop Accomplished 🟩 Pre-Existing Resources ⬛ Continuing Progress

Parsing Solution 2:
MSA Treebank Transduction



🟦 Workshop Accomplished 🟩 Pre-Existing Resources ⬛ Continuing Progress

Parsing Solution 3:
MSA Grammar Transduction



- Dialect -

Probabilistic TAG

الاولاد كتبو الاشعار

Parser

كتبو

الاشعار الاولاد

- MSA -

Treebank

Probabilistic TAG

Tree Transduction

TAG = Tree Adjoining Grammar

🟦 Workshop Accomplished 🟩 Pre-Existing Resources ⬛ Continuing Progress

What We Have Shown

- Baseline: MSA-trained parser on Levantine
  - Baseline: 53.1%
- This work: a small amount of effort improves
  - Small lexicon, 2 syntactic rules: 60.2%
- Comparison: a large amount of effort for treebanking improves more
  - Annotate 11,000 words: 69.3%

Summary: Introduction

- 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
- Want to quickly develop dialect parsers without need for annontation
- Exploit knowledge of differences MSA/dialects to be able to

Global Overview

- Introduction
- **Student Presentation: Safi Shareef**
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion

بسم الله الرحمن الرحيم

Arabic Dialect Text Classification

Student Project Proposal

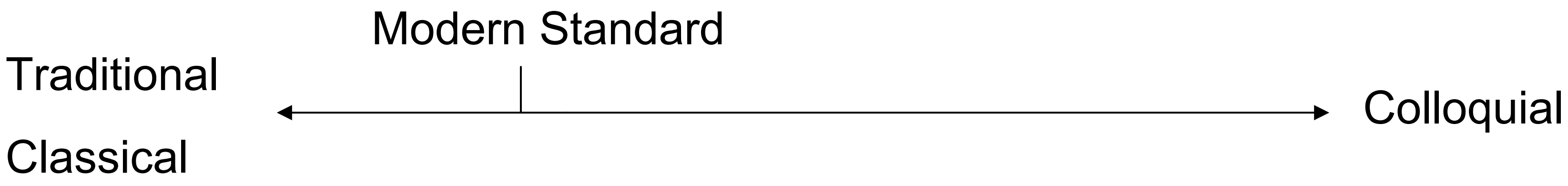
Advisor: Nizar Habash
Student: Safi Shareef

Columbia University, NY
Johns Hopkins University, MD

August 17, 2005

Background

- Arabic Diglossia
  - Standard Arabic: formal, primarily written
  - Arabic Dialects: informal, primarily spoken
  - Differences in phonology, morphology, syntax, lexicon
  - Regional Dialect differences (Iraqi, Egyptian, Levantine, etc.)
- Spectrum of modern Arabic language forms
  - Hints toward content



Code Switching

- MSA & Dialect mixing within the same text

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.

Computational Issues

- Modern Standard Arabic
  - Plethora of resources/applications
  - Textual Corpora
  - Treebanks
  - Morphological Analyzers/Generators
- Arabic Dialects
  - Limited or no resources
  - Many dialects with varying degrees of support

Dialect Detection (Identification)

- Motivation
  - Create more consistent and robust language models
    - Machine translation
      - e.g. Translate into IRQ in colloquial form
  - Application matching
    - What lexicon, analyzer, translation system to use?
    - Dialect ID as additional feature to different applications
      - Information retrieval, information extraction, etc.

Types of Dialect Classification

- Document-based vs. Word-based
- Single Dialect vs. Multiple Dialect
- Form of Dialect

Dimensions of Classification

	Single Dialect	Multiple Dialect
Word	Classify word as MSA or DIA	Classify Word as MSA, IRQ, LEV, EGY, GLF,etc.
Document	Classify document as MSA or DIA, spectrum of Classical ←→Colloquial	Classify Document as MSA, IRQ, LEV, EGY, GLF,etc.

Difficulty of Dialect Identification…

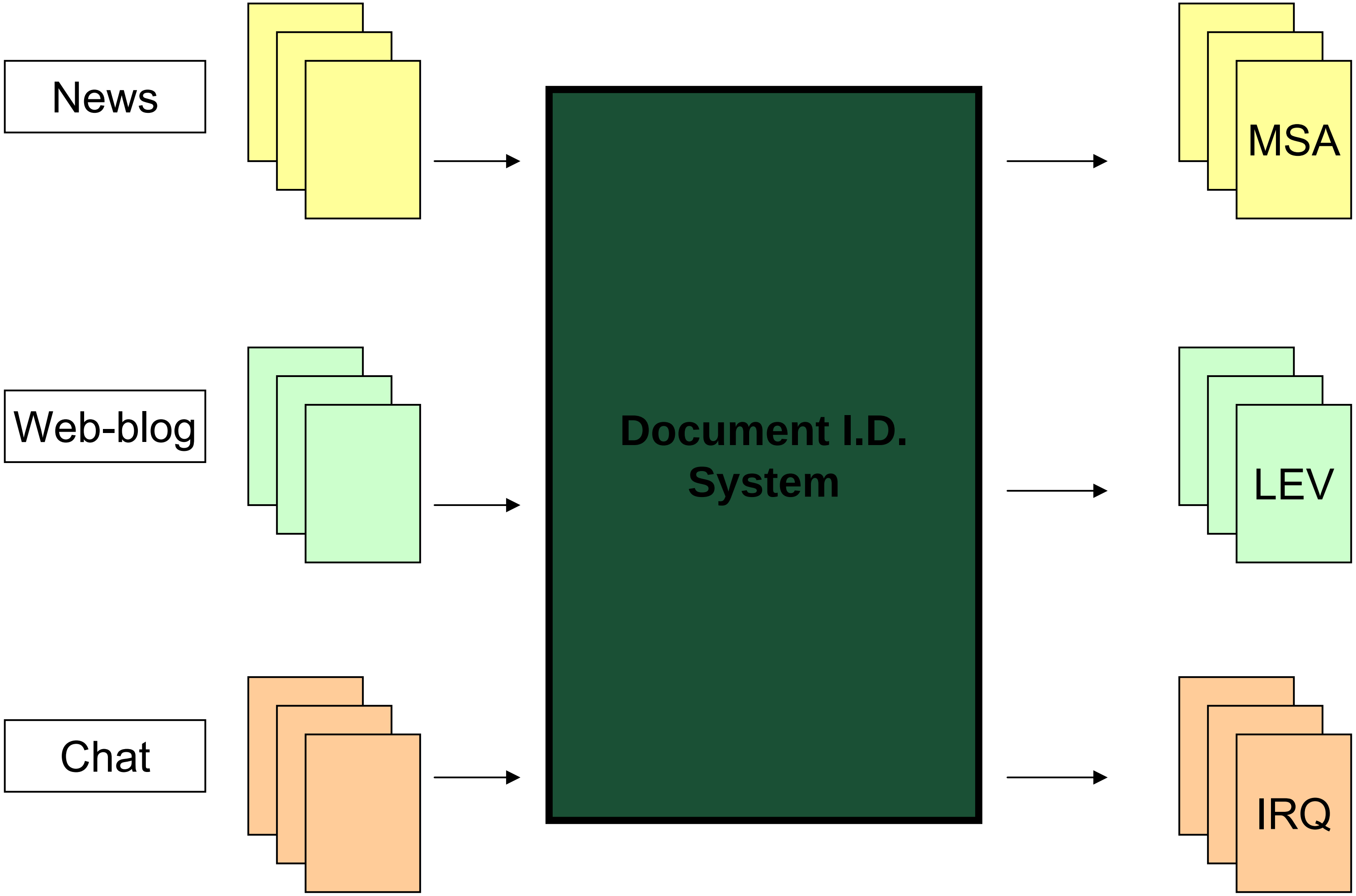
- Research Challenges
  - Require annotated development and test sets
    - Creating annotating resources (i.e. determining dialect)
  - Other resource requirements:
    - e.g. Word analyzers

	Single Dialect	Multiple Dialect
Word	* Hard to annotate * Need resources	* Harder to annotate * Need more resources
Document	URL annotated Corpora Textual resources that originate from known dialectal region	

The Problem Being Addressed…

- Document-level Multiple Dialect Classification
  - No Resources exist to identify an Arabic document’s dialect
    - Unannotated Corpora exists!
      - (e.g. news groups, blogs, interviews, etc.)
  - Encompasses single dialect document-level classification
  - Precursor to word-level classification

Proposal



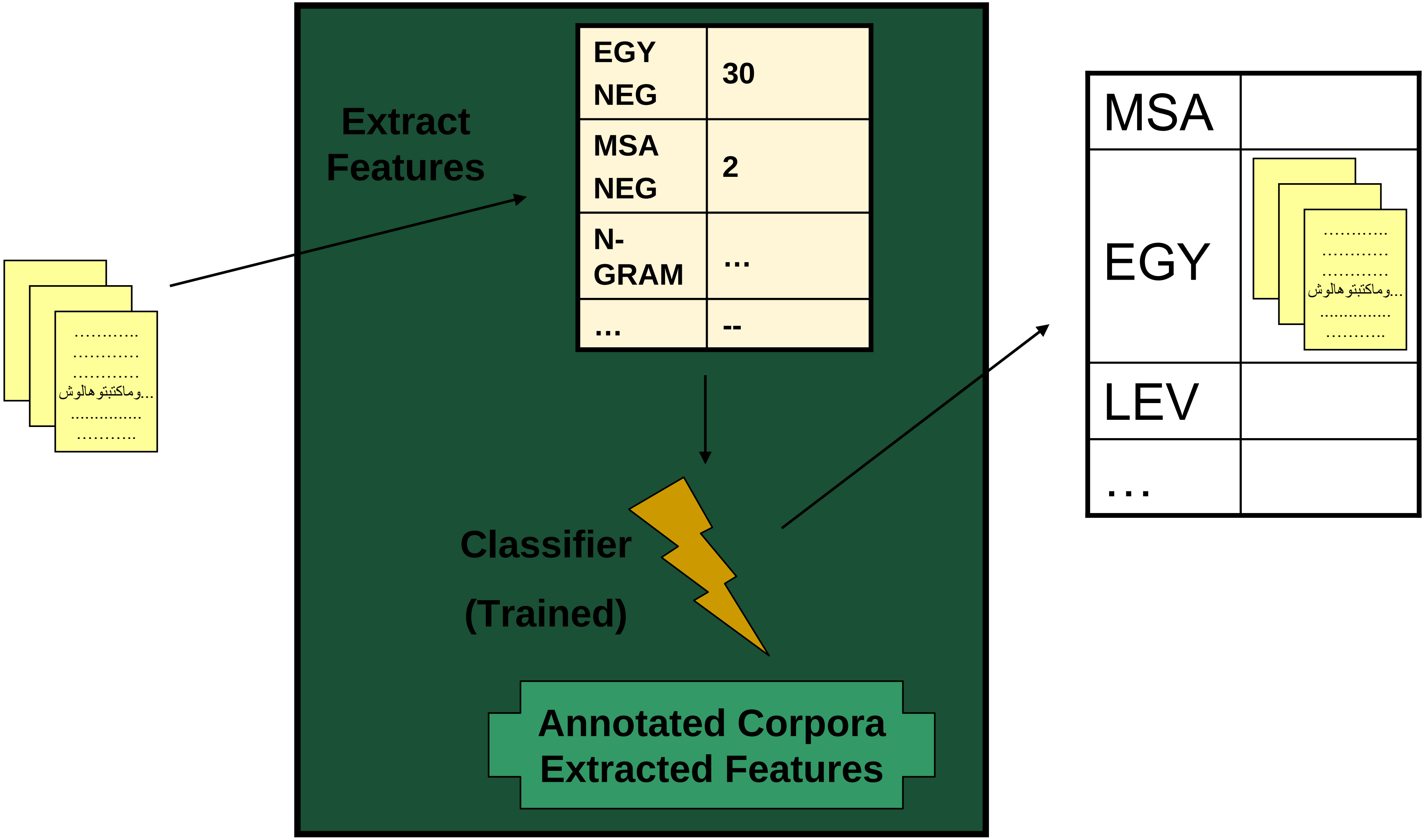
Proposed Solution

- Develop a text level analyzer to rank Arabic text (at the document level) on likelihood of being LEV, EGP, IRQ, MSA, etc …
- Resources
  - Multidialectal corpus annotated by region
    - e.g. use URL of newsgroups
  - Dialect-specific wordlists
  - Any available word-level applications
    - e.g. morphological analyzer

Arabic Dialect Classification vs. Language Identification

- Language Identification
  - Different orthographies
  - Primarily unique vocabulary
- Arabic Dialect Classification
  - Not a simple Text Categorization Problem
    - Same orthography
    - Similar word roots
    - Non-uniform text
      - Code-switching

Proposed Approach



EGY NEG	30
MSA NEG	2
N-GRAM	…
…	--

MSA	
EGY	
LEV	
…	

Global Overview

- Introduction
- Student Presentation: Safi Shareef
- **Student Presentation: Vincent Lacey**
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion

بسم الله الرحمن الرحيم

Statistical Mappings of Multiword Expressions Across Multilingual Corpora

Student Project Proposal

Proposal by

Vincent Lacey	Georgia Tech
Advisor: Mona Diab	Columbia
Sponsor: Chin-Hui Lee	Georgia Tech

First, some motivation:

"Ya be trippin' wit' dat tight truck jewelry."

LEXICON

Ya – Ya; Yes; Okay; You

Be – Be, Are, Is

Trippin' – Tripping, Falling, High, Crazy

Wit' – Wits; With

Dat – That

Tight – Constricting; Cool; Nice

Truck – Truck

Jewelry – Jewelry

Truck Jewelry – Gold Jewelry

Yes be falling wits that constricting truck jewelry.

-5.439

Yes be	0.42
be falling	0.50
falling wits	0.05
wits that	0.22
that constricting	0.35
constricting truck	0.15
truck jewelry	0.03

You be high with that cool gold jewelry.

-2.07

You be	0.40
be high	0.65
high with	0.45
with that	0.92
that cool	0.69
cool gold	0.18
gold jewelry	0.63

You are crazy with that nice gold jewelry.

-1.34

You are	0.89
are crazy	0.70
crazy with	0.51
with that	0.92
that nice	0.72
nice gold	0.25
gold jewelry	0.63

Lexical Issues

- Treebank transduction : MSA->Dialect
- Sentence transduction & grammar transduction:
  Dialect->MSA
- 20% of Levantine words are unrecognized by parsers trained on MSA
- No parallel corpora!

Road Map

- **Some Intuition**
- Mapping Single Words
- Preliminary Results

- Proposal: Mapping Multiword Expressions
  - Approach
  - Advantages & Applications
  - Work Plan

Some Intuition

Optimists play video games, **read** magazines and **listen** to the radio more than do pessimists, while pessimists watch more television…

Read the lyrics, **listen**, download and. . .

Who would **read** or even **listen** to this stuff??

R(read, listen) = 0.72

Lo que me gusta hacer...
LEER
ESCUCHAR
MUSICA **Y**
SALIR

R(leer, y) = 0.65

Hoy, con una computadora y un programa especial, una persona ciega puede acceder a la primera biblioteca virtual en lengua hispana para discapacitados visuales, llamada Tiflolibros, y **leer**--mejor dicho, **escuchar** miles de libros por su cuenta.

R(leer, escuchar) = 0.70

Road Map

- Some Intuition
- **Mapping Single Words**
- **Preliminary Results**

- Proposal: Mapping Multiword Expressions
  - Approach
  - Advantages & Applications
  - Work Plan

Mapping Single Words: Spearman

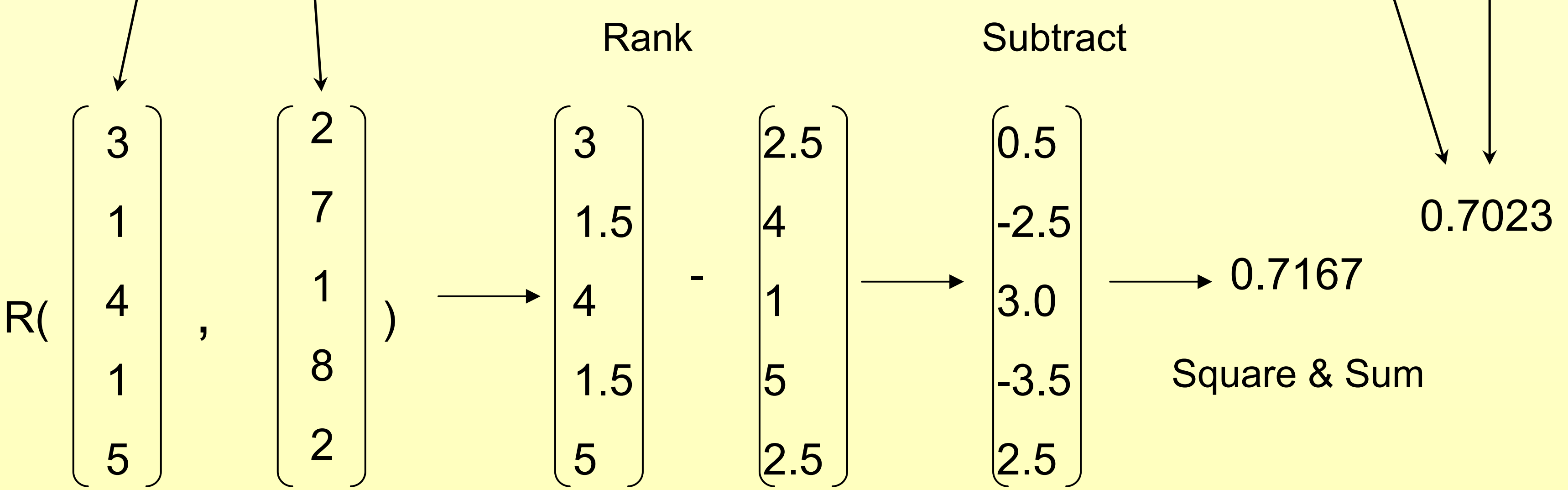
Optimists play video games, **read** magazines and **listen** to the radio more than do pessimists, while pessimists watch more television…

Read the lyrics, **listen**, download and. . .
Who would **read** or even **listen** to this stuff??

$$(R^2) = 1 - \frac{6\sum d^2}{n^3 - n}$$

Lo que me gusta hacer...
LEER ESCUCHAR MUSICA **Y** SALIR

Hoy, con una computadora y un programa especial, una persona ciega puede acceder a la primera biblioteca virtual en lengua hispana para discapacitados visuales, llamada Tiflolibros, y **leer**-- mejor dicho, **escuchar** miles de libros por su cuenta.



Rank

Subtract

$$R\left(\begin{pmatrix}3\\1\\4\\1\\5\end{pmatrix}, \begin{pmatrix}2\\7\\1\\8\\2\end{pmatrix}\right) \rightarrow \begin{pmatrix}3\\1.5\\4\\1.5\\5\end{pmatrix} - \begin{pmatrix}2.5\\4\\1\\5\\2.5\end{pmatrix} \rightarrow \begin{pmatrix}0.5\\-2.5\\3.0\\-3.5\\2.5\end{pmatrix} \rightarrow 0.7167$$

Square & Sum

0.7023

Diab & Finch 2000

Mapping Single Words: Similarity Vectors

Repeat with 3 seed words:

$$\text{truth} = \begin{pmatrix} 0.4305 \\ 0.5547 \\ 0.7120 \end{pmatrix} \quad \text{verisimilitude} = \begin{pmatrix} 0.4326 \\ 0.5937 \\ 0.6785 \end{pmatrix} \quad \text{golden} = \begin{pmatrix} 0.2279 \\ 0.7218 \\ 0.6534 \end{pmatrix}$$

<truth, verisimilitude> = 0.9987

<truth, golden> = 0.9638

Related work: Knight & Koehn 2002

Mapping Single Words: Cognate Filters

Before…. / *After….*

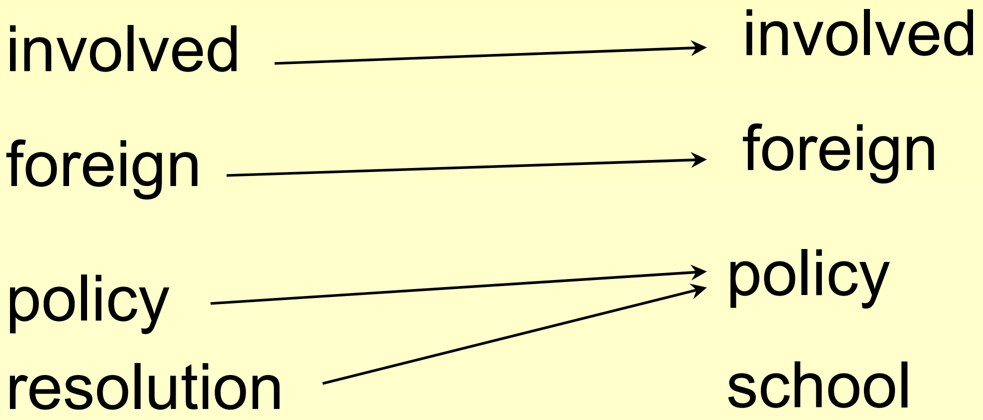
december	degustbe novembebruary octaber december july
family	favestors inflatprices fiaeesiy ppicee inviesators
people	peaople fanvestors inswaters fpeoplyfameirs
china	ohatoasnated japael chisia jassia

$$lcsr = \frac{longest\ common\ substring}{longest\ string}$$

lcsr(government, gouvernement) = 10/12

Melamed 1995

Mapping Single Words: Map Reduction


involved → involved

foreign → foreign

policy → policy

resolution → policy

school

Recall: 25% 50% 100%   Precision: 75% 100%

Preliminary Results: Method Comparison

(English-English comparable corpora)

Methods	Added Entries	Precision
Similarity	1000	86.4%
Similarity+LCSR	1000	92.5%
Similarity+LCSR+MapReduce	**841**	**98.8%**

Preliminary Results: Comparable Corpora Analysis

English-English Corpora	Precision *	
Size (words)	Comparable	Related
100M	99.7% (889)	87.6% (381)
20M	99.2% (825)	84.2% (319)
4M	96.3% (719)	77.7% (286)

Arabic MSA-MSA Corpora	Precision *	
Size (words)	Comparable	Related
100M	99.3% (764)	96.5% (654)
20M	98.2% (756)	87.1% (465)
4M	94.0% (625)	70.9% (288)

Comparable: Same genre (“same” newswire), overlapping coverage time

Related: Same genre (different newswire), some overlapping coverage time

*type precision

Road Map

- Some Intuition
- Mapping Single Words
- Preliminary Results

- **Proposal: Mapping Multiword Expressions**
  - Approach
  - Advantages & Applications
  - Work Plan

Approach: Intersecting Sets

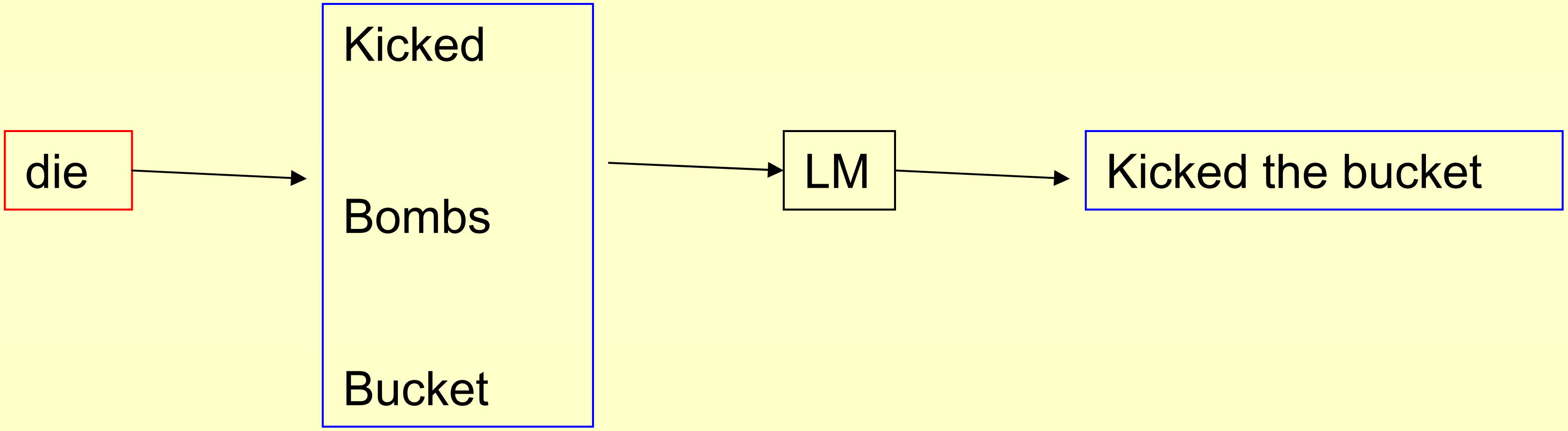
First pass:

kicked	kicked story **die** shove off
the	the of company person **die**
bucket	bucket **die** pail story conclusion

Second pass:

die	passed bombings **bucket** peace **kicked**

Approach: Synthesis



Evaluation

- Using MWE data base at Columbia
- Automated—no human intervention

Advantages & Applications

- No seed lexicon required
- No annotated corpora needed
- *Fast* and extensible

- Word Clustering
- Cross-lingual information retrieval
- **Phrase-based machine translation**

many	many most these other all
issue	issue point ban line force
aid	aid reform power growth investment
ireland	ireland yugoslavia cyprus canada sweden

Work Plan

- Sources: English/Arabic/Chinese Gigaword

- Aug-Sept: Building initial MWE system
- Sept-Oct: Development testing
- Oct-Dec: Final experiments

Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- **Lexicon (Carol Nichols)**
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Grammar Transduction
  - Treebank Transduction
- Conclusion

Local Overview: Lexicon

- Building a lexicon for parsing
  - Get the word to word relations
    - Manual construction
    - Vincent Lacey's presentation (Finch & Diab, 2000)
    - A variant of Rapp (1999)
    - Combination of resources
  - Assign probabilities
- Ways of using lexicons in experiments

Rapp, 1999

English corpus
People who like to read books are interesting.

seed dictionary	
like	ike-lay
books	ooks-bay

Pig latin corpus
e-way ike-lay o-tay ead-ray ooks-bay.

	like	books
are	0	1
read	1	1
…		

	ike-lay	ooks-bay
ead-ray	1	1
e-way	1	0
…		

Automatic Extraction from Comparable Corpora

- Novel extensions to Rapp, 1999:
  - Modification: add best pair to dictionary and iterate
    - When to stop? How “bad” is “bad”?
- English to English corpus: halves of *Emma* by Jane Austen
  - 97% of ~100 words added to dictionary correct
  - 39.5% of other words correct in top candidate
  - 61.5% of other words correct in top 10

Application to LEV-MSA

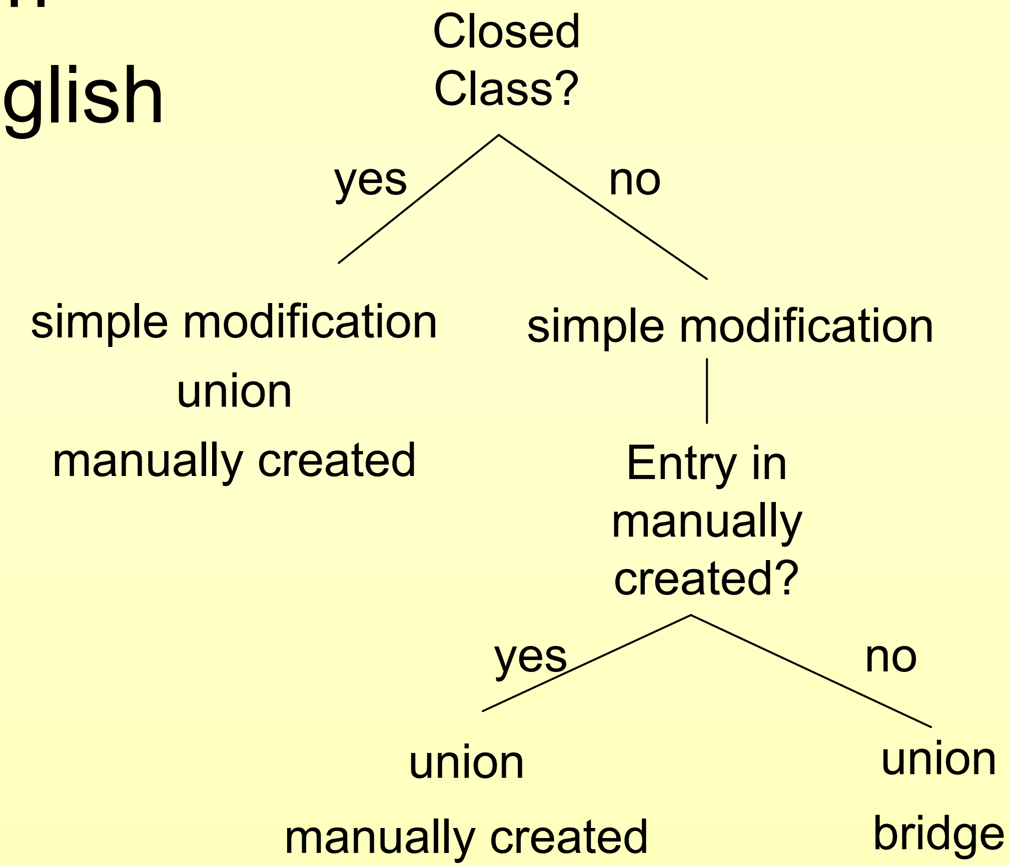
- Levantine development data & part of MSA treebank:
  - Used words that appeared in both corpora as seed dictionary
  - Held out known words: <10% in top 10
  - Manual examination: sometimes clusters on POS
- Explanation:
  - These are small and unrelated corpora
  - If translation is not in other corpus, no chance of finding it!
  - Levantine: speech about family, MSA: text about politics, news

- Contributors: Carol Nichols, Vincent Lacey, Mona Diab, Rebecca Hwa

Manual Construction

- Simple modification
- Bridge through English
- Manually created

- Combination:



Closed Class?

yes

no

simple modification
union
manually created

simple modification

Entry in manually created?

yes

no

union
manually created

union
bridge

Contributors: Nizar Habash

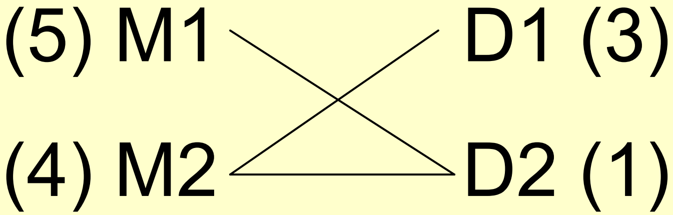
Add Probabilities to Lexicons

- No parallel corpora to compute joint distribution
- Applying EM algorithm using unigram frequency counts from comparable corpora and many-to-many lexicon relations

- Contributors: Khalil Sima’an, Carol Nichols, Rebecca Hwa, Mona Diab, Vincent Lacey

M1 M2
M2
M2 M1
M1
M2 M1
M1

(2) D1
(7) D2



(5) M1

(4) M2

D1 (3)

D2 (1)

(.5) M1
(3.5) M2

D1
D1
D1 D2

Lexicons Used

- Does not rely on corpus specific information
  - Levantine closed class words
  - Top 100 most frequent Levantine words

Small Lexicon

- Uses info from our dev set: occurrence, POS
  - Combined manual lexicon
  - Combined manual lexicon pruned
    - Leaves only non-MSA-like entries and translations found in ATB

Big Lexicon

- Transformed lexemes to surface forms using ARAGEN (Habash, 2004)
- Contributors: Nizar Habash, Carol Nichols, Vincent Lacey

Experiment Variations

POS tags	No Lexicon	Small Lexicon	Big Lexicon
None			
Automatic			
Gold			

Lexical Issues Summary

- Main conclusions:
  - Automatic extraction from comparable corpora for Levantine and MSA is difficult
  - Using small and big lexicons can improve POS tagging and parsing
- Future directions:
  - Try other automatic methods (Ex: tf/idf)
  - Try to find more comparable corpora

Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- **Part-of-Speech Tagging (Rebecca Hwa)**
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Grammar Transduction
  - Treebank Transduction
- Conclusion

POS Tagging

- Assign parts-of-speech to Levantine words

tEny VBP l+ IN +y PRP AlEA}lp NN tmArp NNP ? PUNC

- Correctly tagged input gives higher parsing accuracies

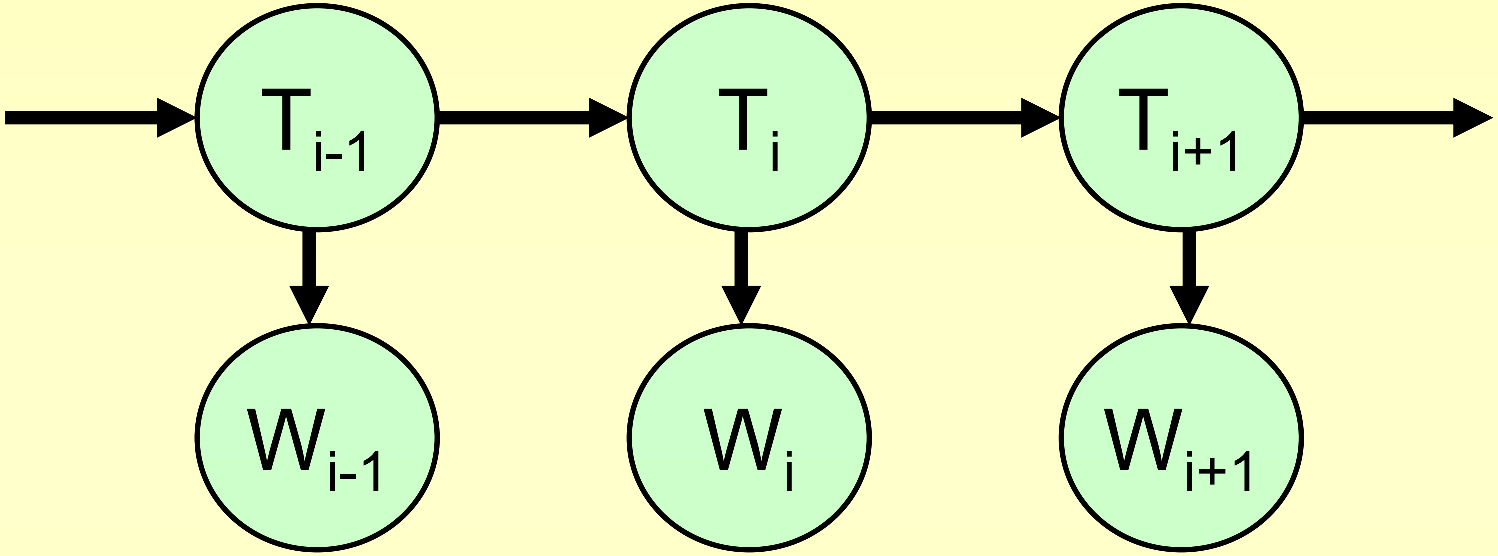
- Assumptions
  - Have MSA resources
  - Levantine data is tokenized
  - Use reduced "Bies" tagset

Contributors: Rebecca Hwa and Roger Levy

Porting from MSA to LEV

- Lexical coverage challenge
  - 80% of word tokens overlap
  - 60% of word types overlap
  - 6% of the overlapped types (10% of tokens) have different tags
- Approaches
  - Exploit readily available resources
  - Augment model to reflect characteristics of the language

Basic Tagging Model: HMM



- Transition distributions: $P(T_i | T_{i-1})$
- Emission distributions: $P(W_i | T_i)$
- Initial model: MSA Bigram
  - Trained on 587K manually tagged MSA words

Tagging LEV with MSA Model

- Baselines: Train on MSA
  - Test on MSA: 93.4%
  - Test on LEV:
    - Dev (11.1K words): 68.8%
    - Test (10.6K words): 64.4%
- Train on LEV
  - 10-fold cross validation on LEV Dev: 82.9%
  - Train on LEV Dev, Test on LEV test: 80.2%
- Higher accuracies (~70%) are possible with models such as SVM (Diab et al., 2004)

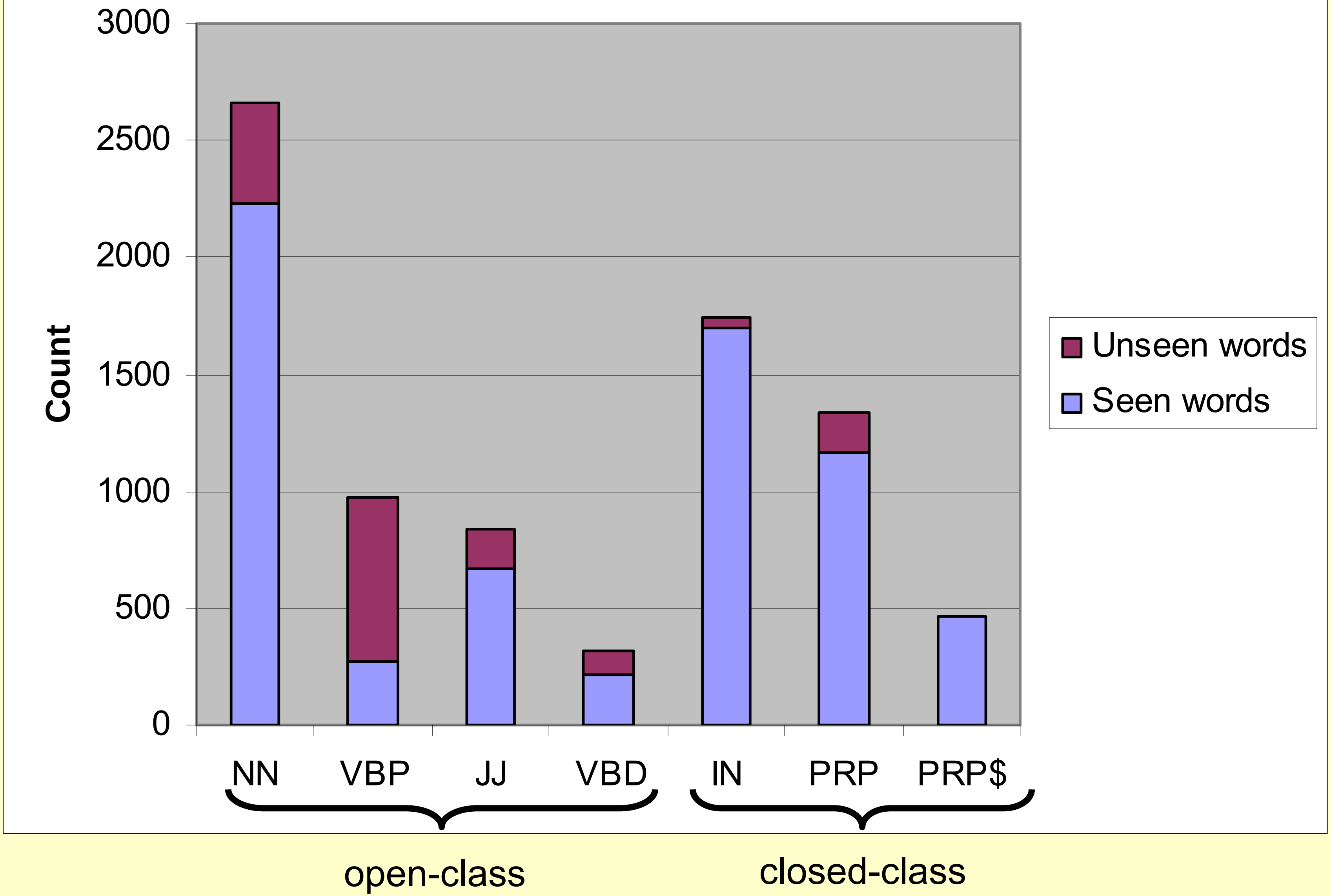
Naïve Porting

- Assume no change in transitions $P(T_i|T_{i-1})$
- Adapt emission probabilities $P(W|T)$
  - Reclaim mass from MSA-only words
  - Redistribute mass to LEV-only words proportional to unigram frequency
- Unsupervised re-training with EM
- Results on LEV dev:
  - 70.2% without retraining
  - 70.7% after one iteration of EM
  - Further retraining hurts performance
- Result on LEV test: 66.1%

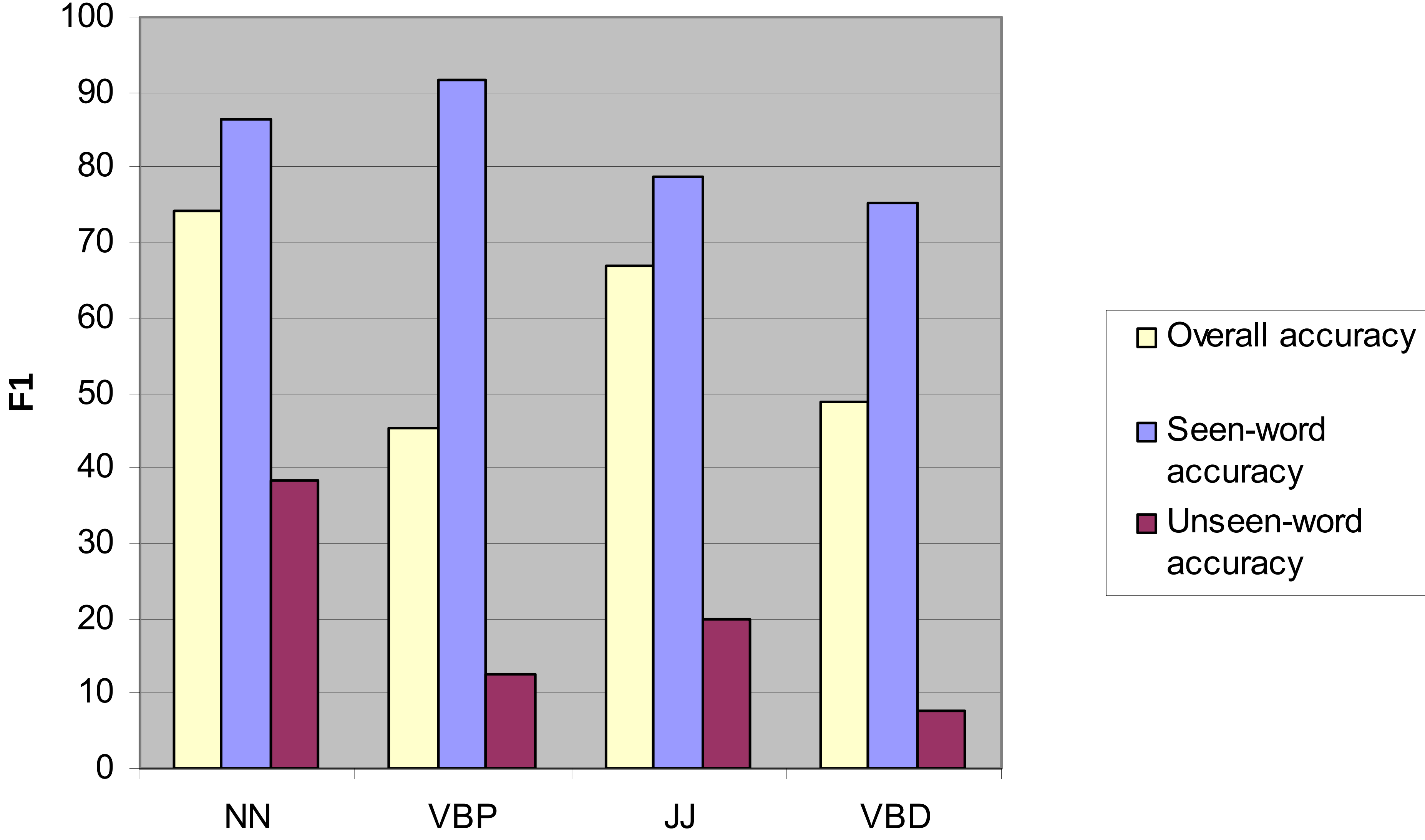
Error Analyses on LEV Dev

- Transition
  - Genre/Domain differences affect transition probabilities
  - Retraining transition probabilities improves accuracy
- Emission
  - Accuracy of MSA-LEV shared words: 84.4%
  - Accuracy of LEV-only words: 16.9%
  - Frequent errors on closed-class words
- Retraining
  - Naïve porting doesn’t give EM enough constraints

Relative proportion of seen/unseen words in Levantine development set



Tagging accuracy for open-class parts of speech



Exploit Resources

- Minimal linguistic knowledge
  - Closed-class vs. open-class
  - Gather stats on initial and final two letters
    - e.g., *Al*+ suggests Noun, Adj.
  - Most words have one or two possible Bies tags
- Translation lexicons
  - “Small” vs. “Big”
- Tagged dialect sentences
- Morphological analyzer (Duh&Kirchhoff, 2005)

Tagging Results on LEV Test

POS tags	No Lexicon	Small Lexicon	Big Lexicon
None			
Automatic			
Gold			

Tagging Results on LEV Test

	No Lexicon	Small Lexicon	Big Lexicon
Naive Port	66.6%		
Minimal Linguistic Knowledge	70.5%	77.0%	78.2%
+100 Tagged LEV Sentences (300 words)	78.3%	79.9%	79.3%

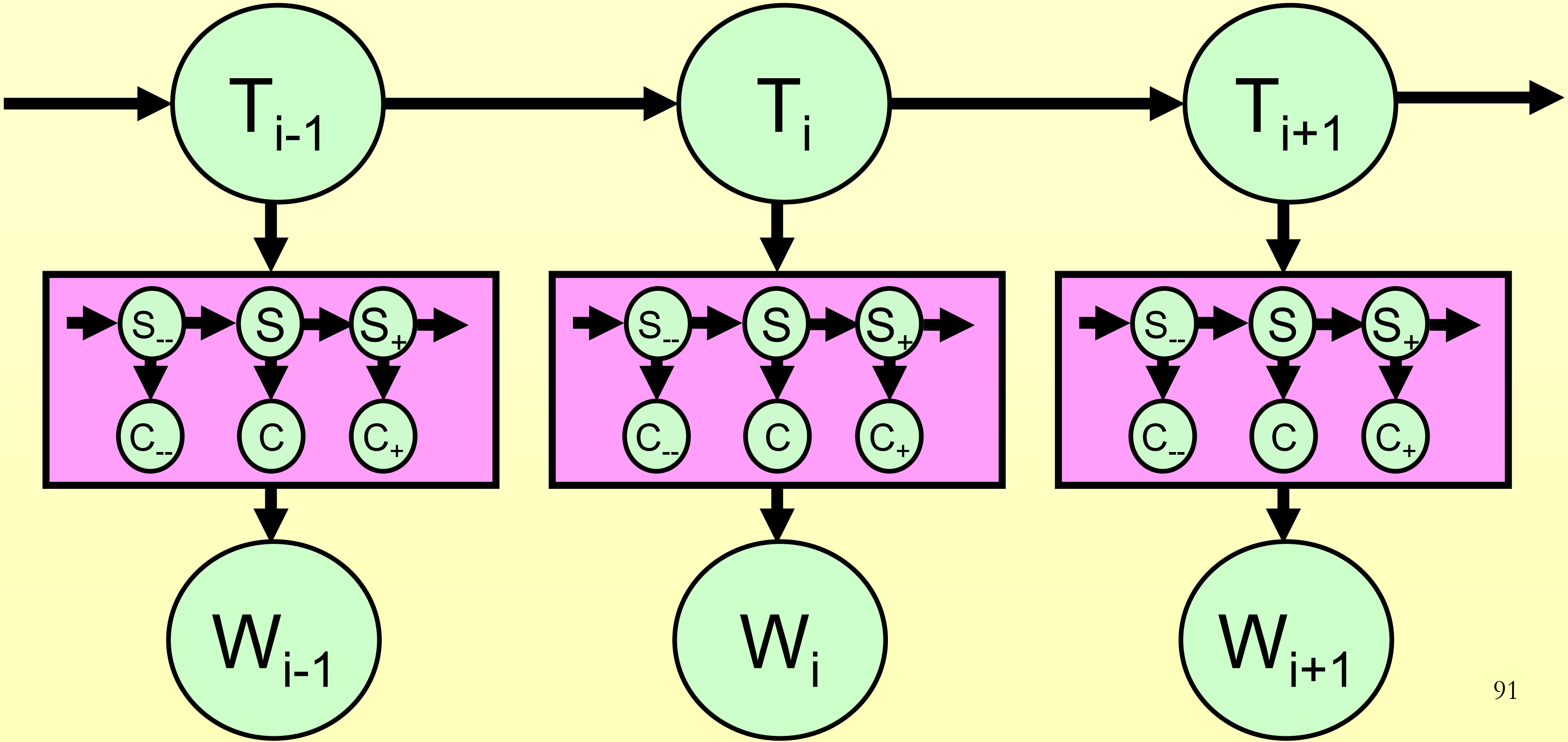
- Baseline: MSA as-is: 64.4%
- Supervised (~11K tagged LEV words): 80.2%

Ongoing Work: Augment Tagging Model

- Distributional methods promising for POS
  - Clark 2000, 2003: completely unsupervised
- We have much more distr. information
  - Some MSA parameters are useful
- LEV words' *internal* structure constrainable
  - morphological regularities useful for POS clustering (Clark 2003)

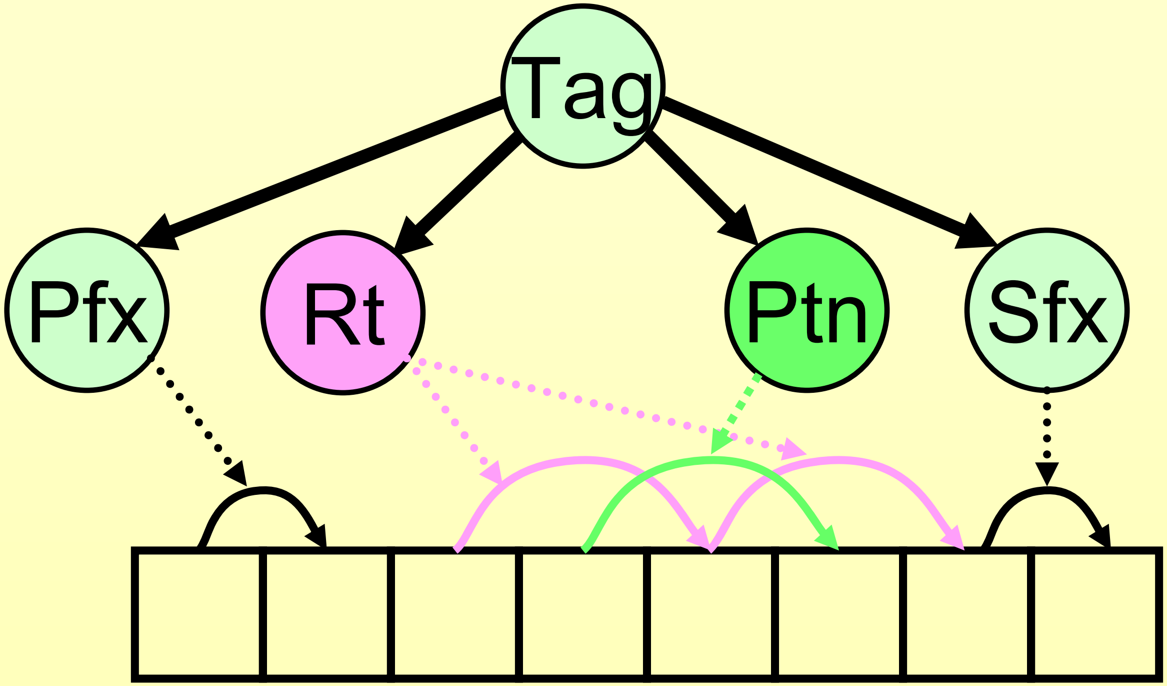
Version 1: Simple Morphology

- P(W|T) determined with character HMM
  - each POS has separate char. HMM



Version 2: Root-Template Morphology

- Character HMM doesn't capture lots of Arabic morphological structure
- Templates determine open-class POS



POS Tagging Summary

- Lexical coverage is a major challenge
- Linguistic knowledge helps
- Translation lexicons are useful resources
  - Small lexicon offers biggest bang for $$
- Ongoing work: improve model to take advantage of morphological features

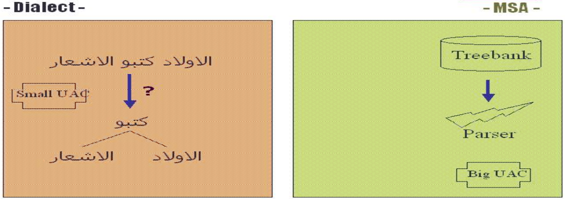
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - **Introduction and Baselines (Khalil Sima’an)**
  - Sentence Transduction
  - Treebank Transduction
  - Grammar Transduction
- Conclusion

Parsing Arabic Dialect

Baselines for Parsing

Parsing Arabic Dialects:
The Problem



- Dialect -

الاولاد كتبو الاشعار

Small UAC

?

كتبو

الاشعار الاولاد

- MSA -

Treebank

Parser

Big UAC

Baselines for Parsing LEV

Alternative baseline approaches to parsing Levantine:

- **Unsupervised:** Unsupervised induction
- **MSA-supervised:** Train statistical parser on MSA treebank

Hypothetical:

- **Treebanking:** Train on small LEV treebank (13k words)

Our approach:

- **Without treebanking:** Porting MSA parsers to LEV
  Exploring simple word transduction

Reminder: LEV Data

MSA is Newswire text – LEV is Callhome

For this project, the following strictly speech phenomena were removed from the LEV data (M. Diab):

- EDITED (restarts) and INTJ (interjections)
- PRN (Parentheticals) and UNFINISHED constituents
- All *resulting* SINGLETON trees

Resulting data:

- *Dev-set* (1928 sentences) and *Test-set* (2051 sentences)
- Average sentence length: about 5.5 wds/sen.

Reported results are F1 scores.

Baselines: Unsupervised Parsers for LEV

Unsupervised induction by PCFG [Klein & Manning].

Induce structure for the gold POS tagged LEV dev-set (R. Levy):

Model	**Unlab Brack.**	**Lab Brack.**	**Untyped Dep.**	**Typed Dep.**
Unsupervised	42.6	–	50.9	–

Baselines: MSA Parsers for LEV (1)

MSA Treebank PCFG (R. Levy and K. Sima'an).

Model	**Unlab Brack.**	**Lab Brack.**	**Untyped Dep.**	**Typed Dep.**
TB PCFG(Free)	63.5	**50.5**	56.1	34.7
TB PCFG(+Gold)	71.7	**60.4**	66.1	49.0
TB PCFG(+Smooth)	73.0	**62.3**	66.2	51.6

Most improvement (10%) comes from gold tagging!

Free: bare words input
+Gold: gold POS tagged input
+Smooth: (+Gold) + smoothed model

Baselines: MSA Parsers for LEV (2)

Gold tagged input:

Model	**Unlab Brack.**	**Lab Brack.**	**Untyped Dep.**	**Typed Dep.**
TB PCFG (+G+S)	73.0	**62.3**	66.2	51.6
Blex.dep. (Bikel)[1]		**60.9**		
Treegram (Sima'an)	73.7	**62.9**	68.7	51.5
STAG (Chiang)	73.6	**63.0**	71.0	52.8

Free POS Tags

STAG (Chiang)		**55.3**		

Treebank PCFG doing as well as lexicalized parsers?

[1] Gold POS tags partially enforced (N. Habash).

Treebanking LEV: A Reference Point

NOTE: This serves only as a reference point!

Train a statistical parser on 13k words LEV treebank.
How good a LEV parser will we have?

D. Chiang:

- Ten-fold split LEV-dev-set (90%/10%) train/test sets
- Trained STAG-parser on train, tested on test:

Free tags: F1 = **67.7**   Gold tags: F1 = **72.6**

Questions:

- Will injecting LEV knowledge into MSA parsers give more?
- What kind of knowledge? How hard is it to come by?

Some Numbers About Lexical Differences

Without morphological normalization on either side.

In the LEV dev-set:

- 21% of word tokens are not in MSA treebank
- 27% of $\langle word, tag \rangle$ occurrences are not in MSA treebank

The Three Fundamental Approaches

Sentence: Translate LEV sentences to MSA sentences

Treebank: Translate MSA treebank into LEV

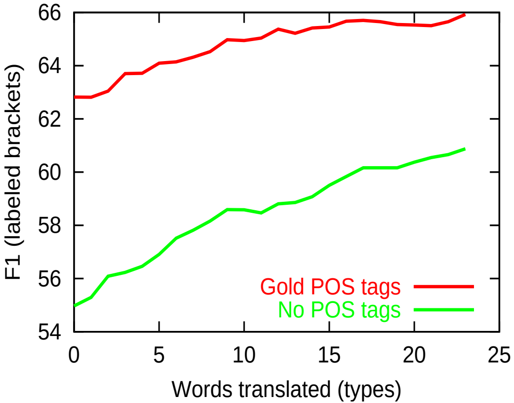
Grammar: Translate prob. MSA grammar into LEV grammar

Common to all three approaches: word-to-word translation

Let us try simple word-to-word translation

A Cheap Extension to the Baseline

Hypothesis: translating a small number of words will improve parsing accuracy significantly (D. Chiang & N. Habash).



Simple transduction "half-way" to LEV treebank parser

Preview of Baseline Results

Model	**Unlab Brack.**	**Lab Brack.**	**Untyped Dep.**	**Typed Dep.**
Gold POS Tagged Input				
STAG (Chiang)	73.6	**63.0**	71.0	52.8
Not Tagged Input (Free)				
STAG (Chiang)		**55.3**		

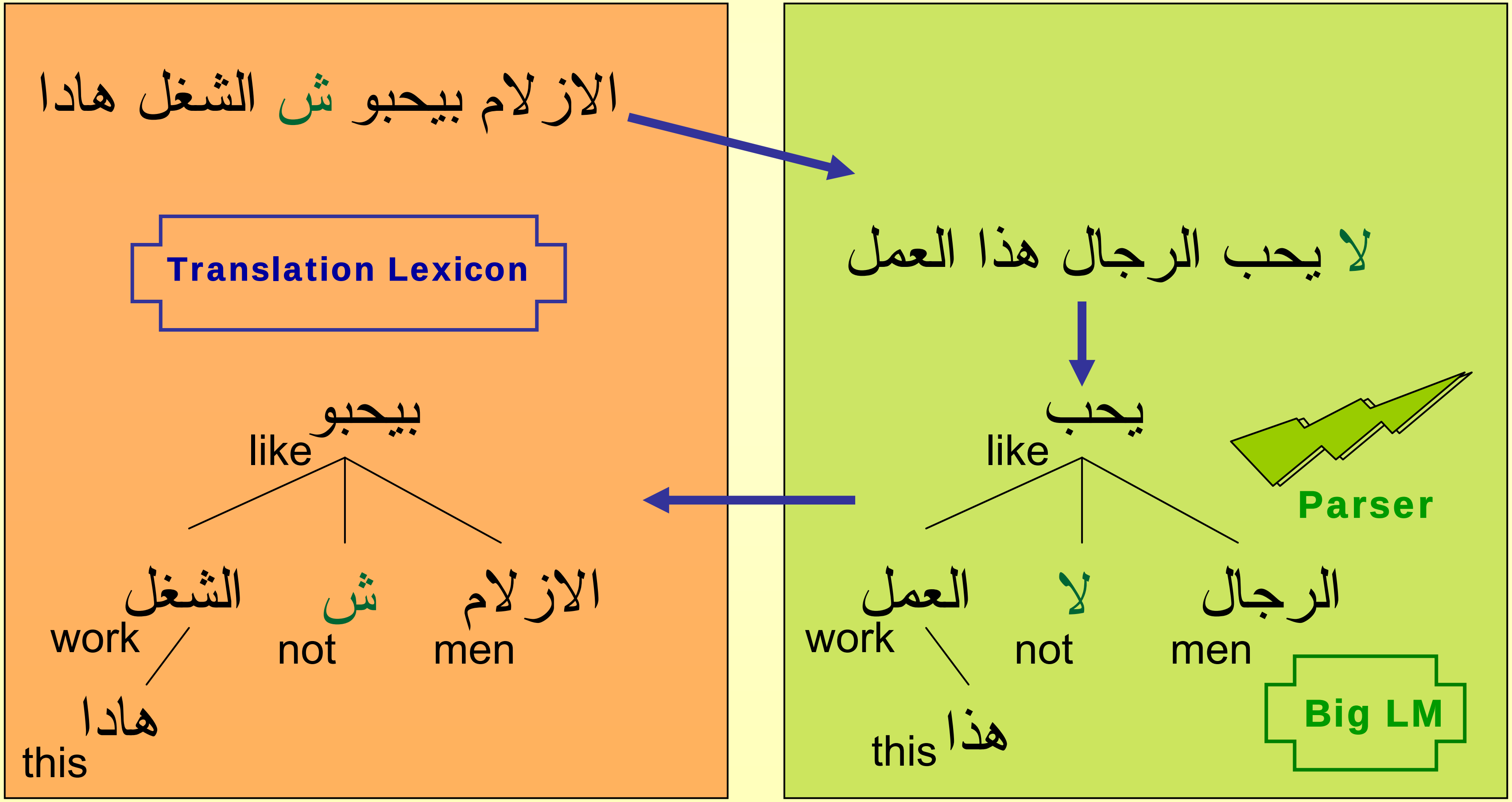
Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
    - Introduction and Baselines
    - **Sentence Transduction (Nizar Habash)**
    - Treebank Transduction
    - Grammar Transduction
- Conclusion

Sentence Transduction Approach

- Dialect -

- MSA -



الازلام بيحبو ش الشغل هادا

Translation Lexicon

بيحبو
like

الشغل
work

ش
not

الازلام
men

هادا
this

لا يحب الرجال هذا العمل

يحب
like

العمل
work

لا
not

الرجال
men

هذا
this

Parser

Big LM

Contributors: Nizar Habash, Safi Shareef, Khalil Sima'an

Intuition/Insight

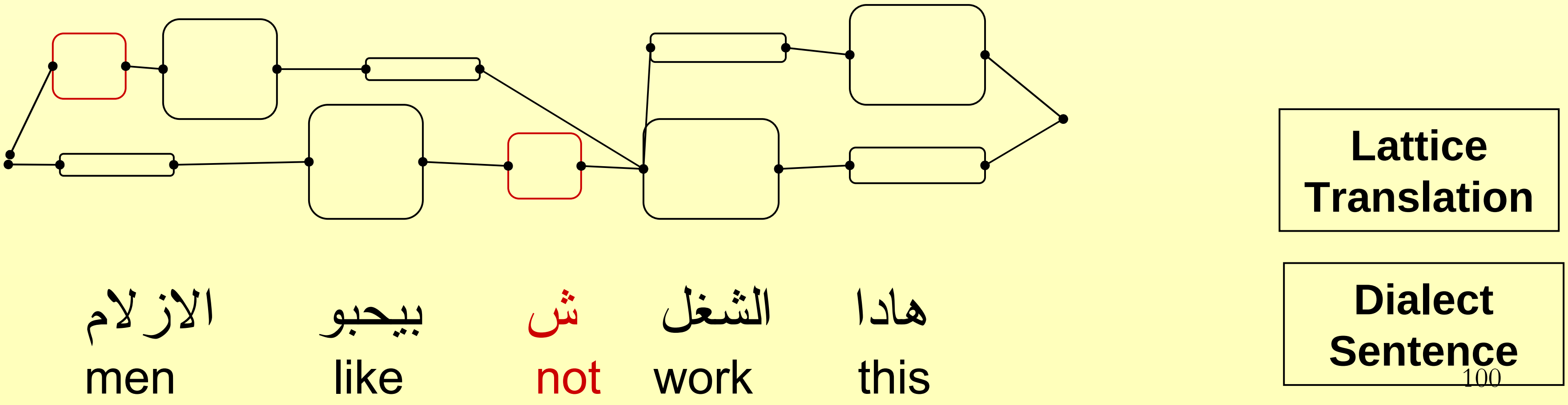
- Translation between closely related languages (MSA/Dialect) is relatively easy compared to translation between unrelated languages (MSA,Dialect/English)
- Dialect-MSA translation is easier than MSA-Dialect translation due to rich MSA resources
  - Surface MSA language models
  - Structural MSA language models
  - MSA grammars

Sentence Transduction Approach

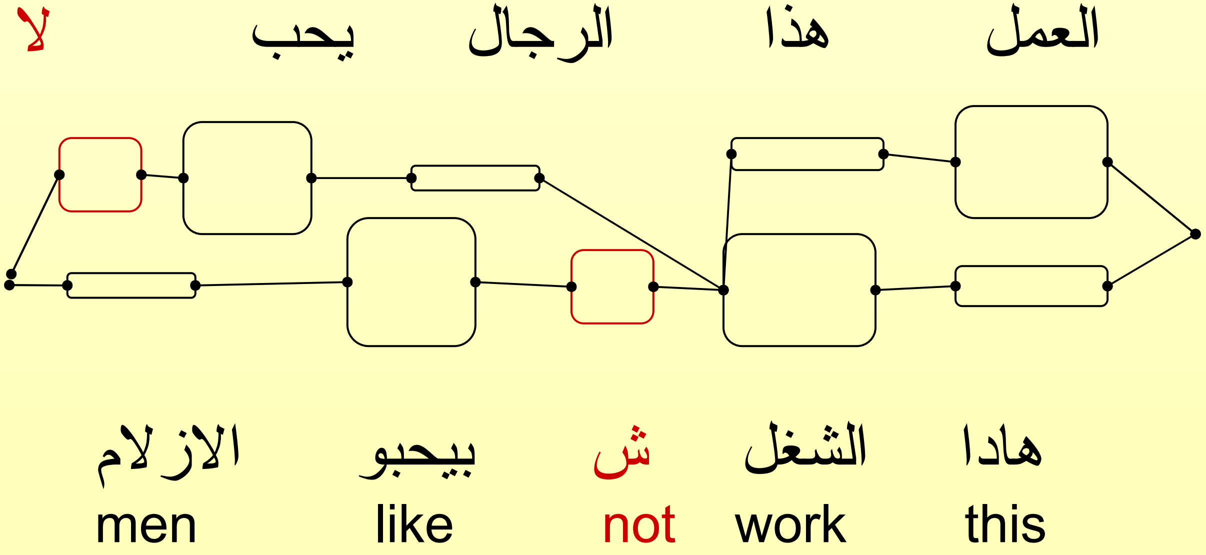
- Advantages
    - MSA translation created as a side product

- Disadvantages
    - No access to structural information for translation
    - Translation can add more ambiguity for parsing
        - Dialect distinct words can become ambiguous MSA words
            - LEV مين myn ‘who’/ من mn ‘from’
            - MSA من mn ‘who/from’

- Translate dialect sentence to MSA lattice
  - Lexical choice under-specified
  - Linear permutations using string matching transformative rules



- Language modeling
  - Select best path in lattice

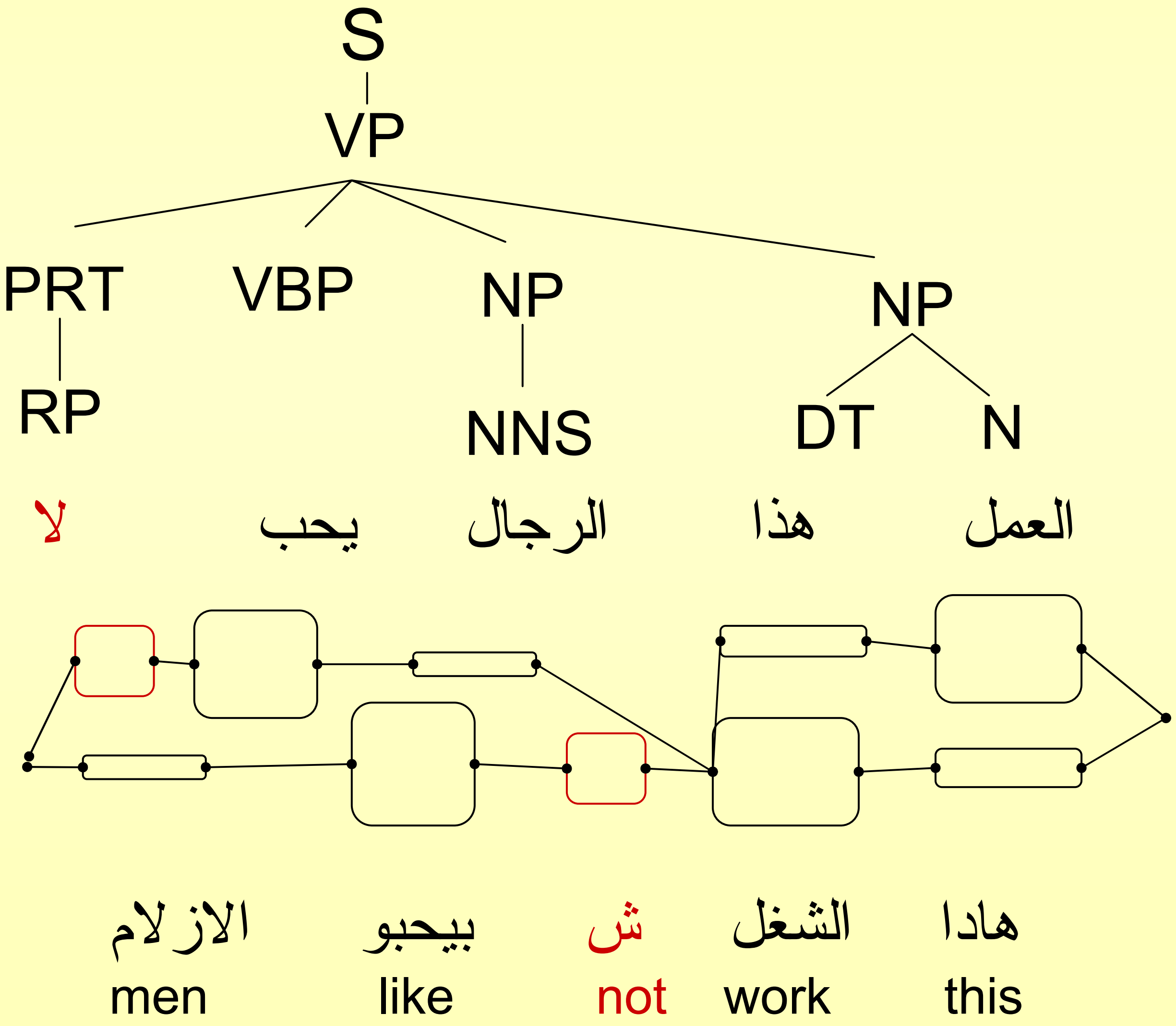


Language Model

Lattice Translation

Dialect Sentence

- MSA Parsing
  - Constituency representation



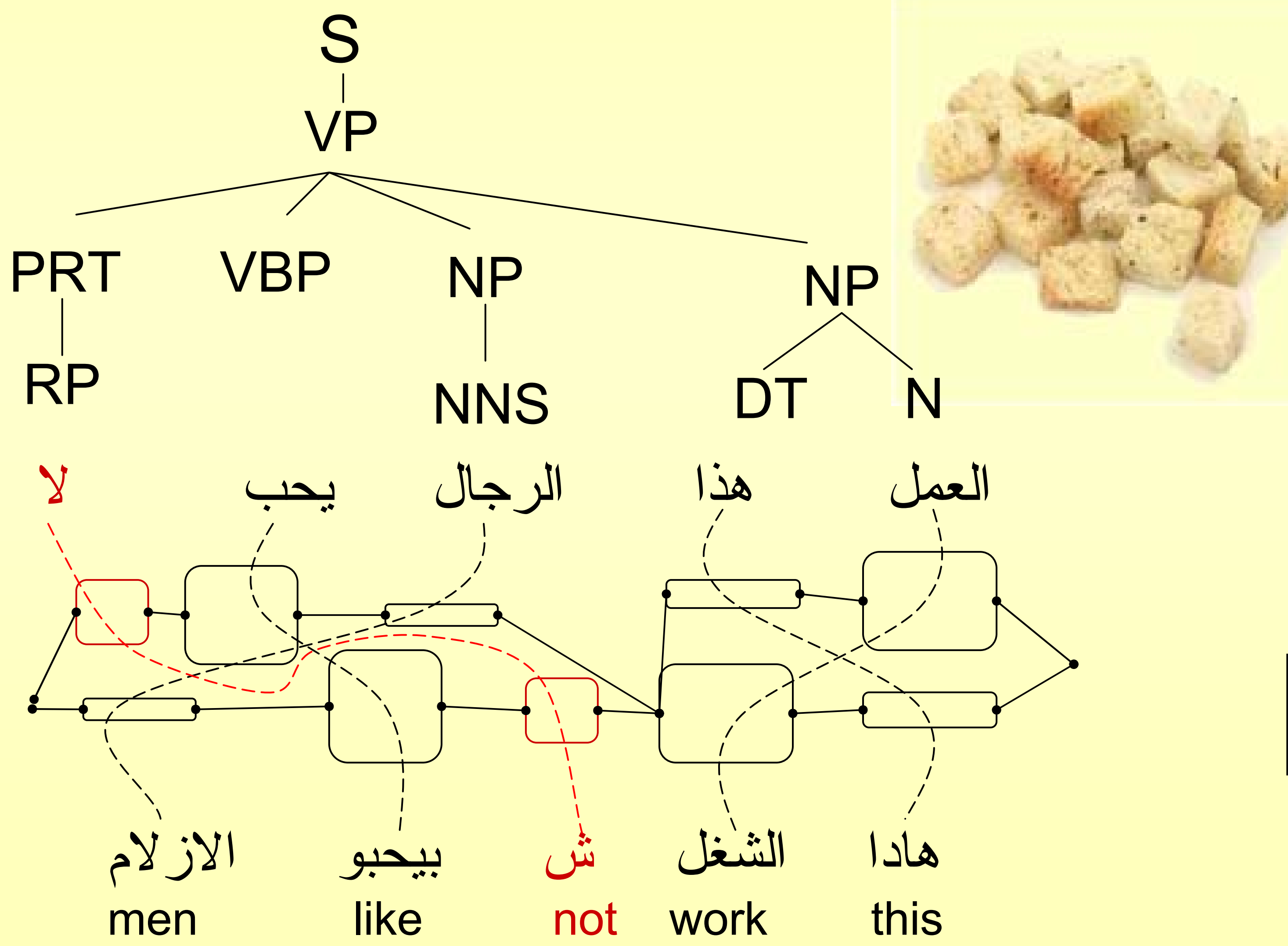
MSA Parsing

Language Model

Lattice Translation

Dialect Sentence

- All along, pass links for dialect word to MSA words



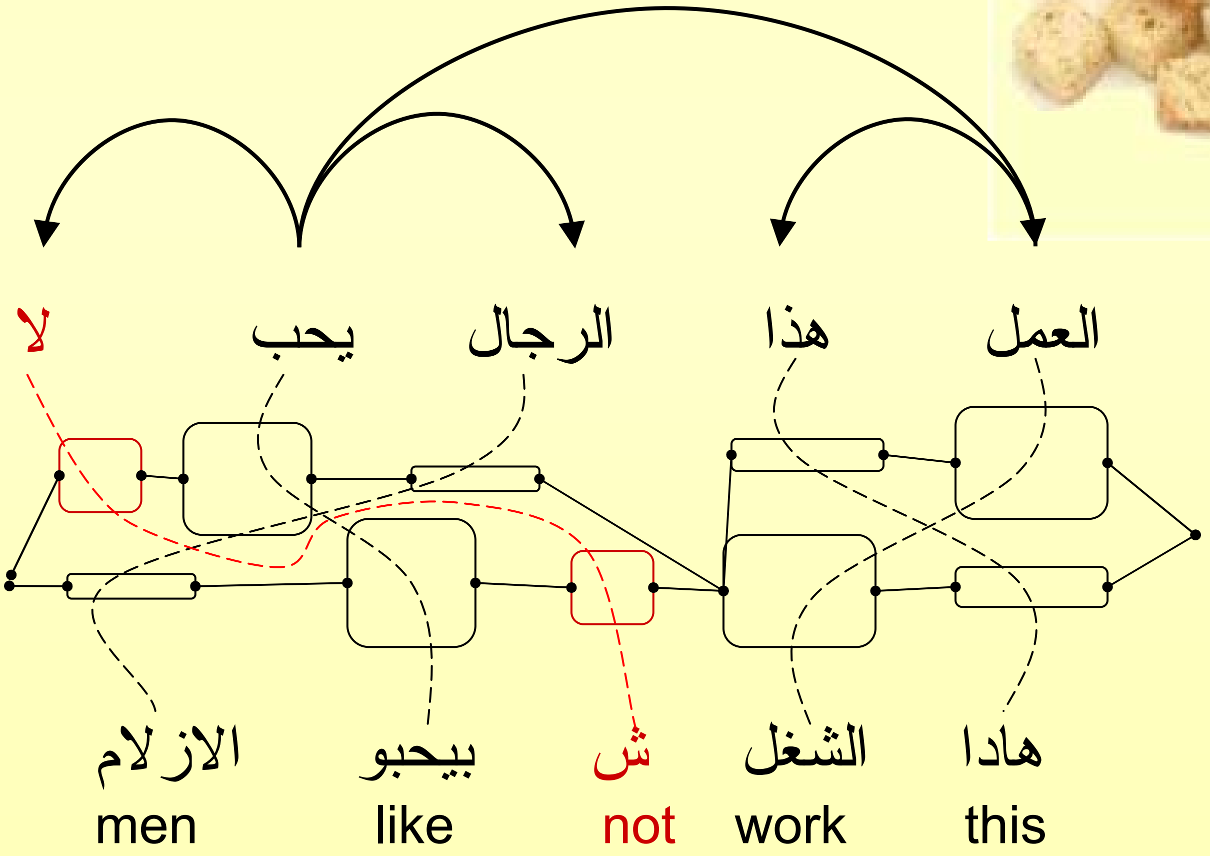
MSA Parsing

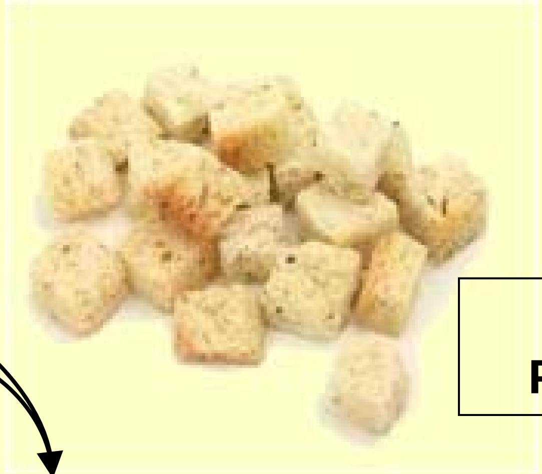
Language Model

Lattice Translation

Dialect Sentence

- Retrace to link dialect words to parse
  - Dependency representation necessary





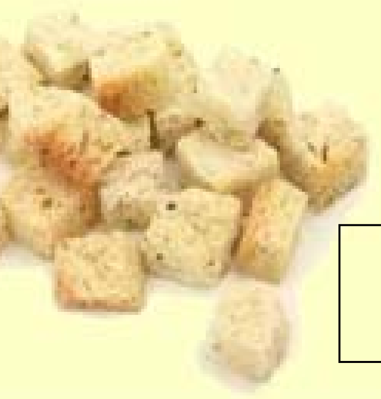
MSA Parsing

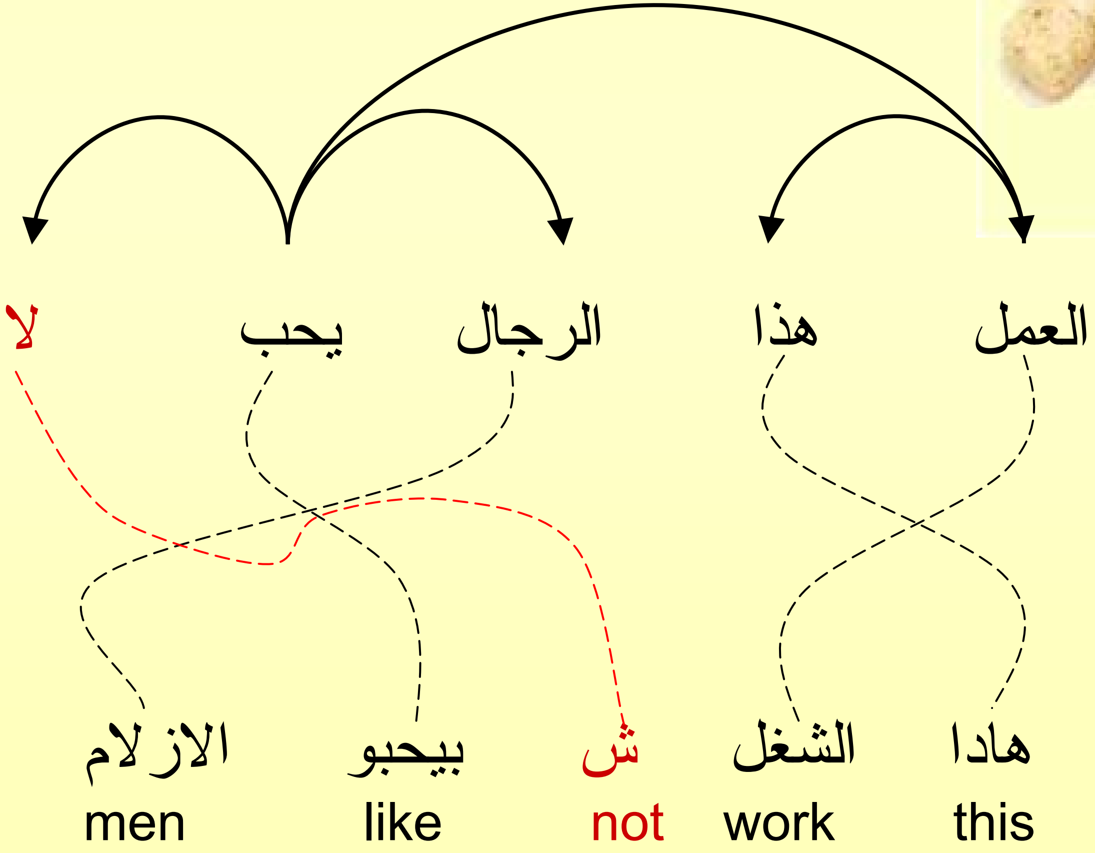
Language Model

Lattice Translation

Dialect Sentence

- Retrace to link dialect words to parse
  - Dependency representation necessary





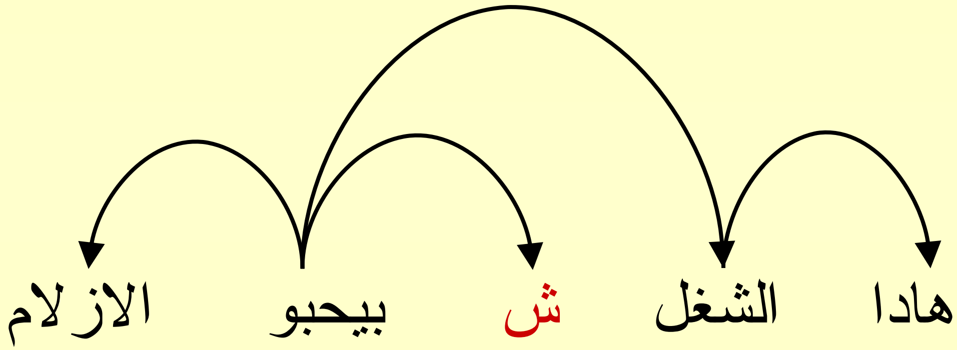
MSA Parsing

Language Model

Lattice Translation

Dialect Sentence

- Retrace to link dialect words to parse
  - Dependency representation necessary



MSA Parsing

Language Model

Lattice Translation

Dialect Sentence

الازلام	بيحبو	ش	الشغل	هادا
men	like	not	work	this

DEV Results

- Bikel Parser, unforced gold tags, uniform translation probabilities
  - PARSEVAL P/R/F1

Tags	No Lexicon	Small Lexicon	Big Lexicon
None	59.4/51.9/55.4	63.8/58.3/61.0	65.3/61.1/63.1
Gold	64.0/58.3/61.0	67.5/63.4/65.3	66.8/63.2/65.0

  - POS tagging accuracy

Tags	No Lexicon	Small Lexicon	Big Lexicon
None	71.3	80.4	83.9
Gold	87.5	91.3	88.6

TEST vs DEV

- PARSEVAL P/R/F1

	Lexicon None		**Lexicon** Small	
Tags	DEV	TEST	DEV	TEST
None	55.4	53.5	61.0	57.7
Gold	61.0	60.2	65.3	64.0

- POS tagging accuracy

	Lexicon None		**Lexicon** Small	
Tags	DEV	TEST	DEV	TEST
None	71.3	67.4	80.4	74.6
Gold	87.5	86.6	91.3	89.8

Additional Experiments

- EM translation probabilities
  - Not much or consistently helpful
- Lattice Parsing alternative (Khalil Sima'an)
  - Using a structural LM (but no additional surface LM)
  - No EM probs used
  - PARSEVAL F1 score

	Lexicon None		**Lexicon** Small	
Tags	DEV	TEST	DEV	TEST
Gold	62.9	62.0	63.0	61.9

Linear Permutation Experiment

- Negation permutation
  - **V $/RP → lA/RP V**
- 3% in Dev, 2% in Test
- Dependency accuracy

	Lexicon Small			
	DEV		TEST	
Tags	NoPerm	PermNeg	NoPerm	PermNeg
Gold	69.6	69.7	67.6	67.3

Conclusions & Future Plans

- Framework for sentence transduction approach
- 22% reduction on pos tagging error (DEV=32%)
- 9% reduction on F1 labeled constituent error (DEV=13%)

- Explore a larger space of permutations
- Better LMs on MSA
- Integrate surface LM probabilities in lattice parsing approach
- Use Treebank/Grammar transduction parses (without lexical translation)

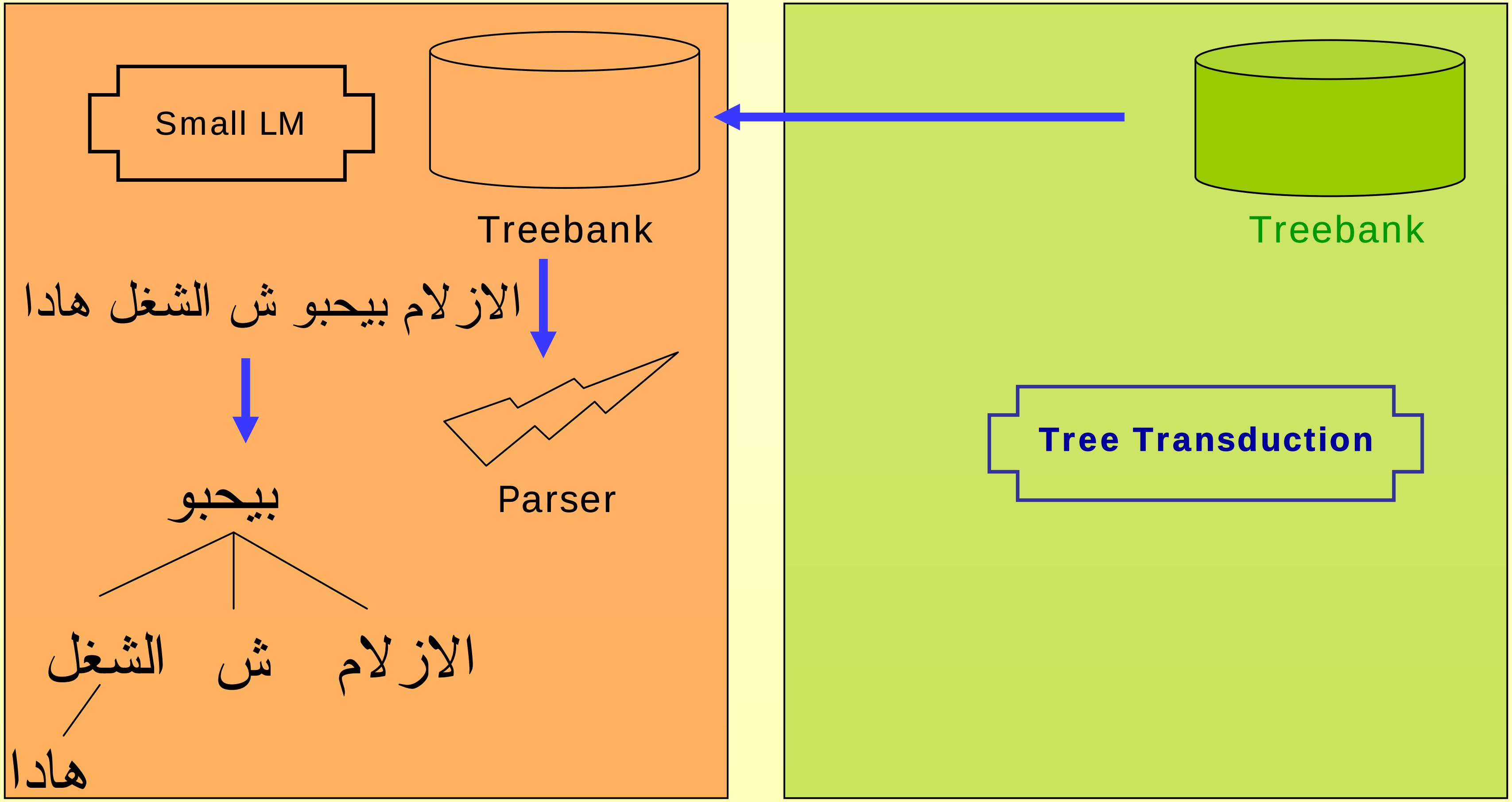
Global Overview

- Introduction (Owen Rambow)
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - **Treebank Transduction (Mona Diab)**
  - Grammar Transduction
- Conclusion

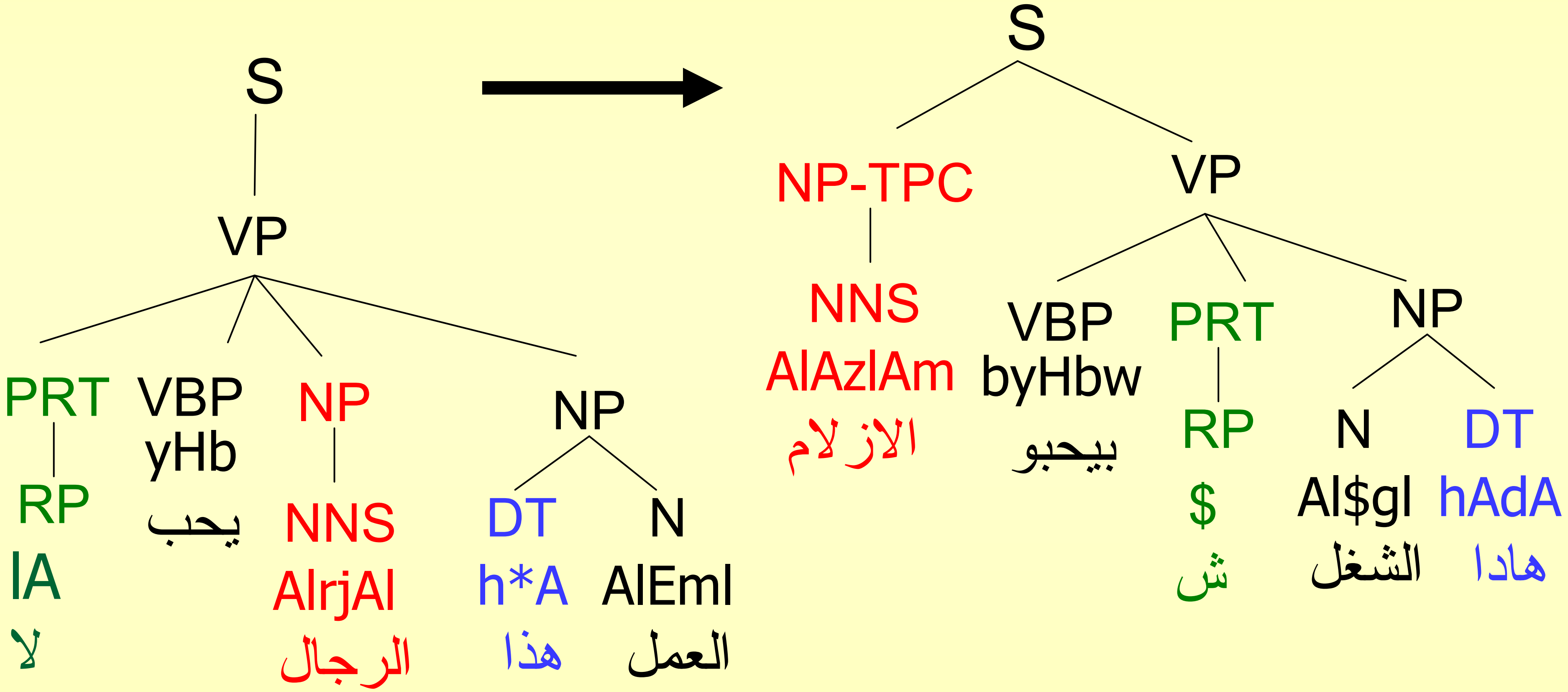
MSA Treebank Transduction

- Dialect -

- MSA -



Objective


S
VP
PRT
RP
lA
لا
VBP
yHb
يحب
NP
NNS
AlrjAl
الرجال
NP
DT
h*A
هذا
N
AlEml
العمل
S
NP-TPC
NNS
AlAzlAm
الازلام
VP
VBP
byHbw
بيحبو
PRT
RP
$
ش
NP
N
Al$gl
الشغل
DT
hAdA
هادا

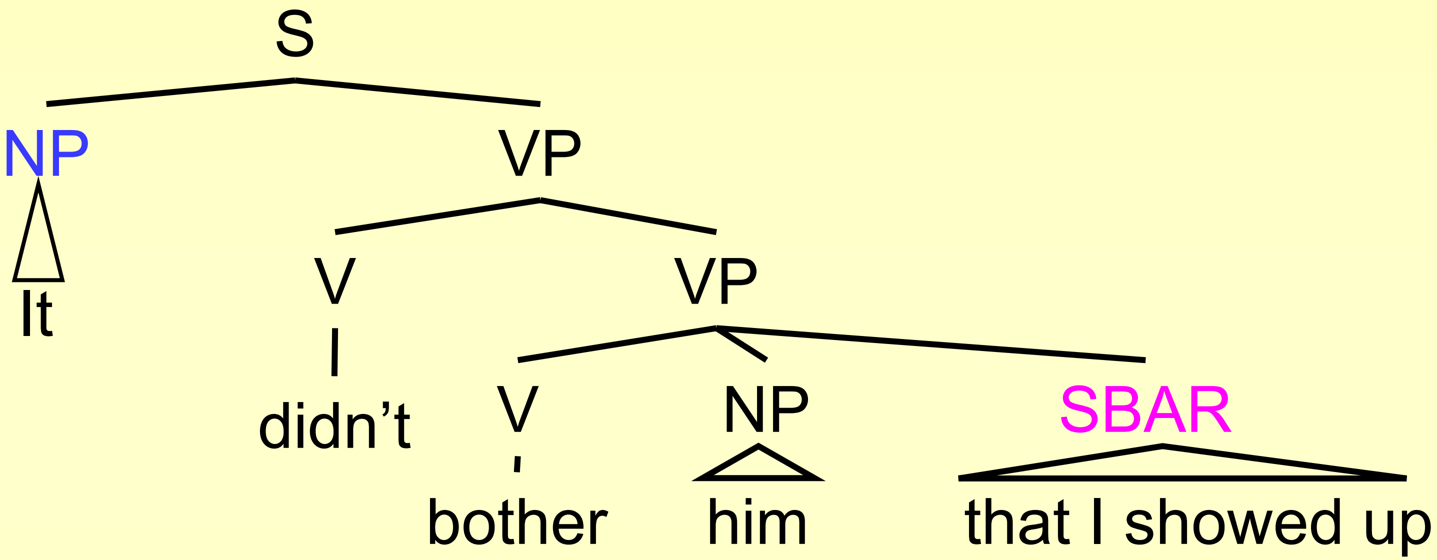
Approach

- Structural Manipulations
  - Tree normalizations
  - Syntactic transformations
- Lexical Manipulations
  - Lexical translations
  - Morphological transformations

Resources Required

- MSA Treebank (provided by LDC)
- Knowledge of systematic structural transformations (scholar seeded knowledge)
- Tool to manipulate existing structures (Tregex & Tsurgeon)
- Lexicon of correspondences from MSA to LEV (automatic + hand crafted)
- Evaluation corpus

Tregex (Roger Levy)



descendent through VP chain

headed by

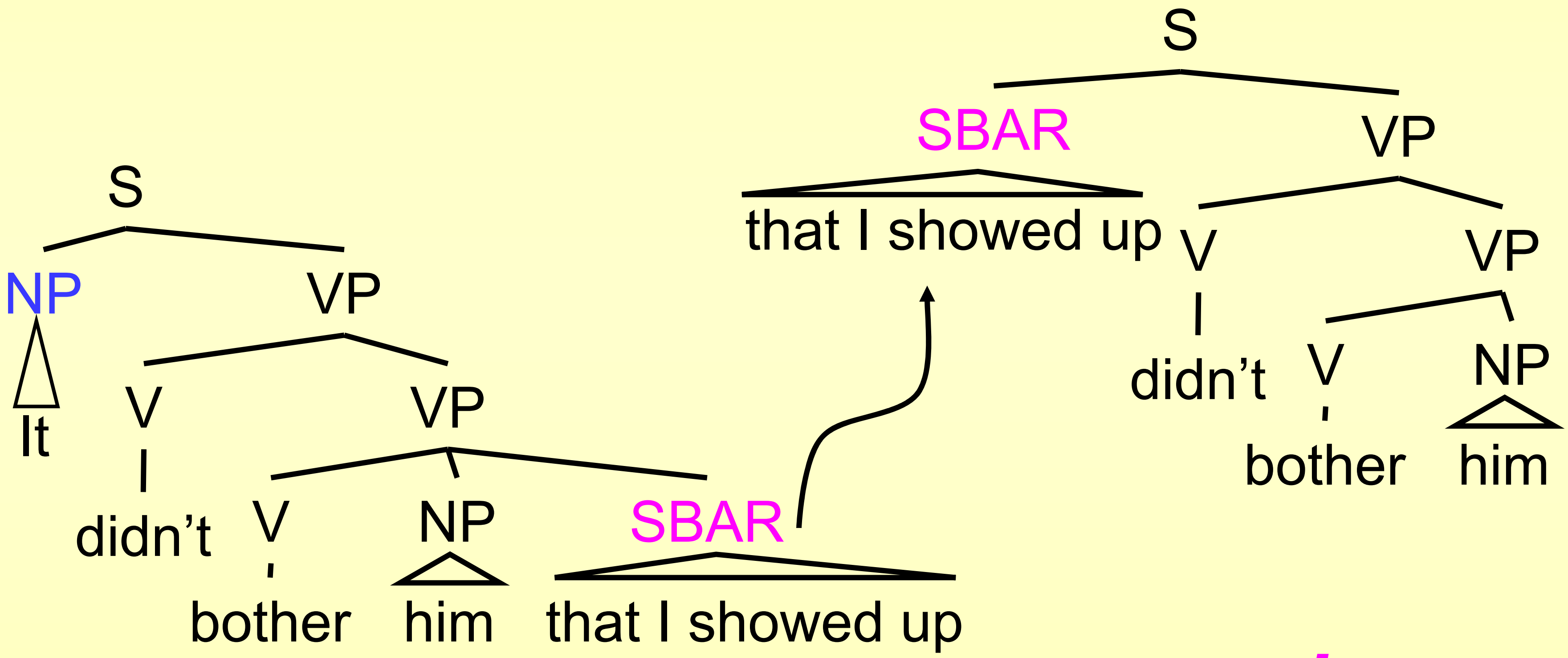
SBAR=sbar > (VP >+VP (S < (NP=np <<# /^[Ii]t$/)))

child-of

dominates

regex “it” or “It”

Tsurgeon (Roger Levy)

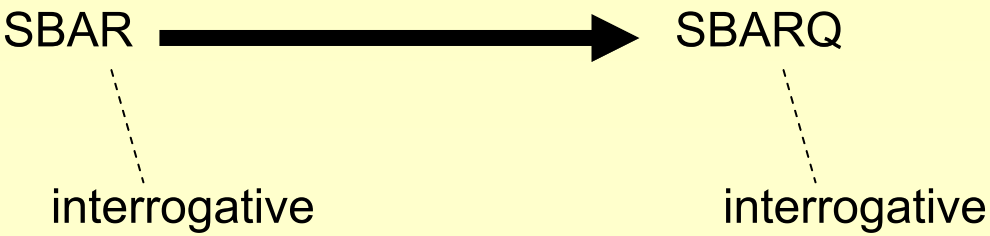


```
prune sbar
replace np sbar
```

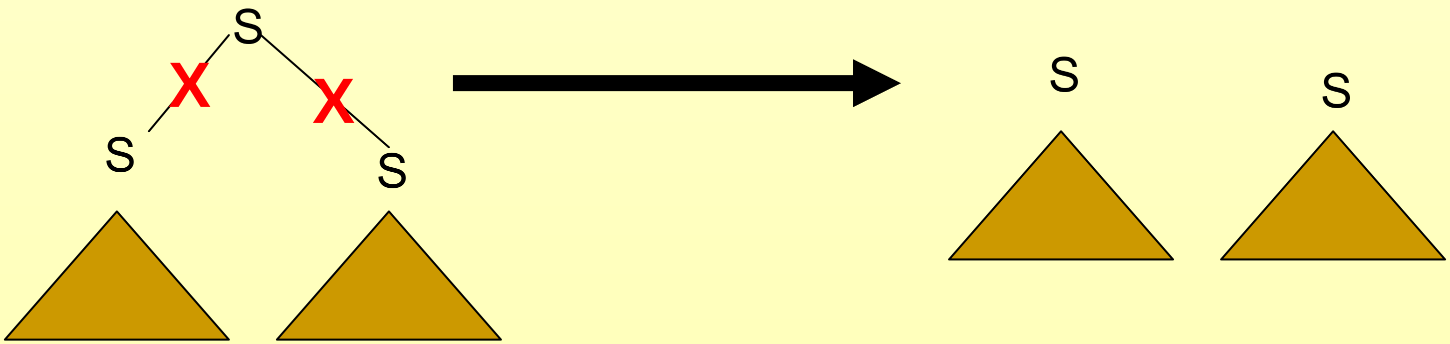
```
SBAR=sbar > (VP >+VP (S < (NP=np <<# /^[Ii]t$/)))
```

Tree Normalizations

Fixing annotation inconsistencies in MSA TB



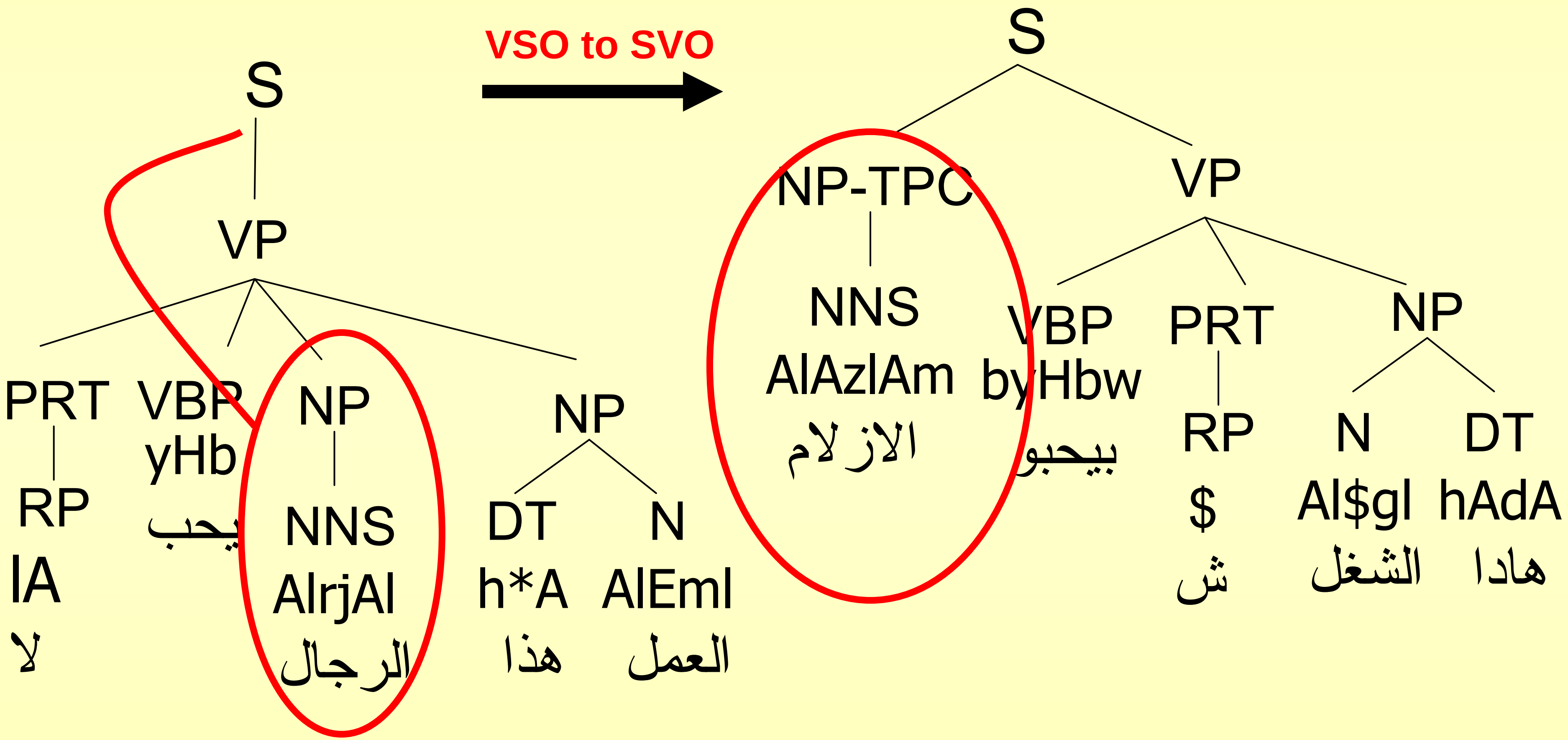
Removing superfluous Ss



Syntactic Transformations

- SVO-VSO
- Fragmentation
- Negation
- Demonstative Pronoun flipping

Syntactic Transformations

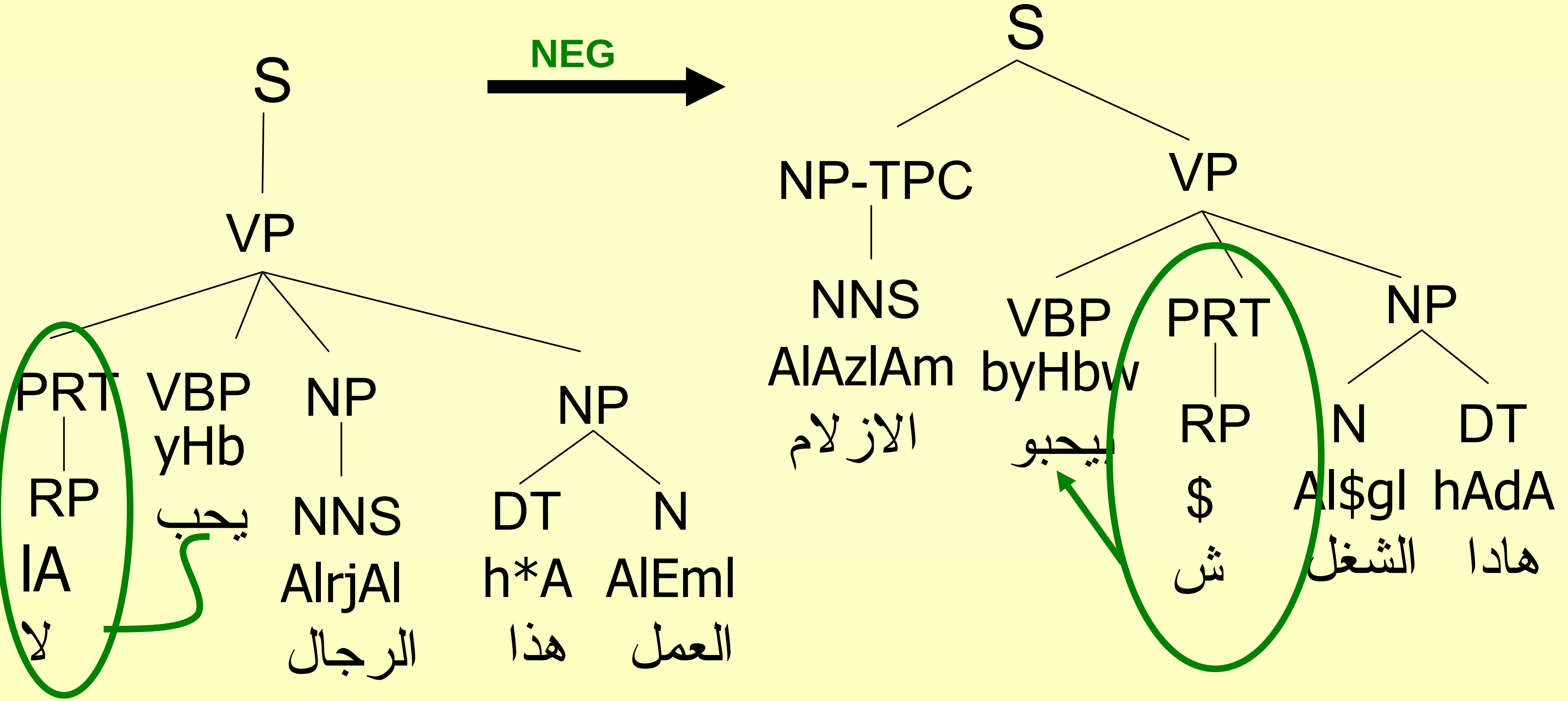


VSO to SVO

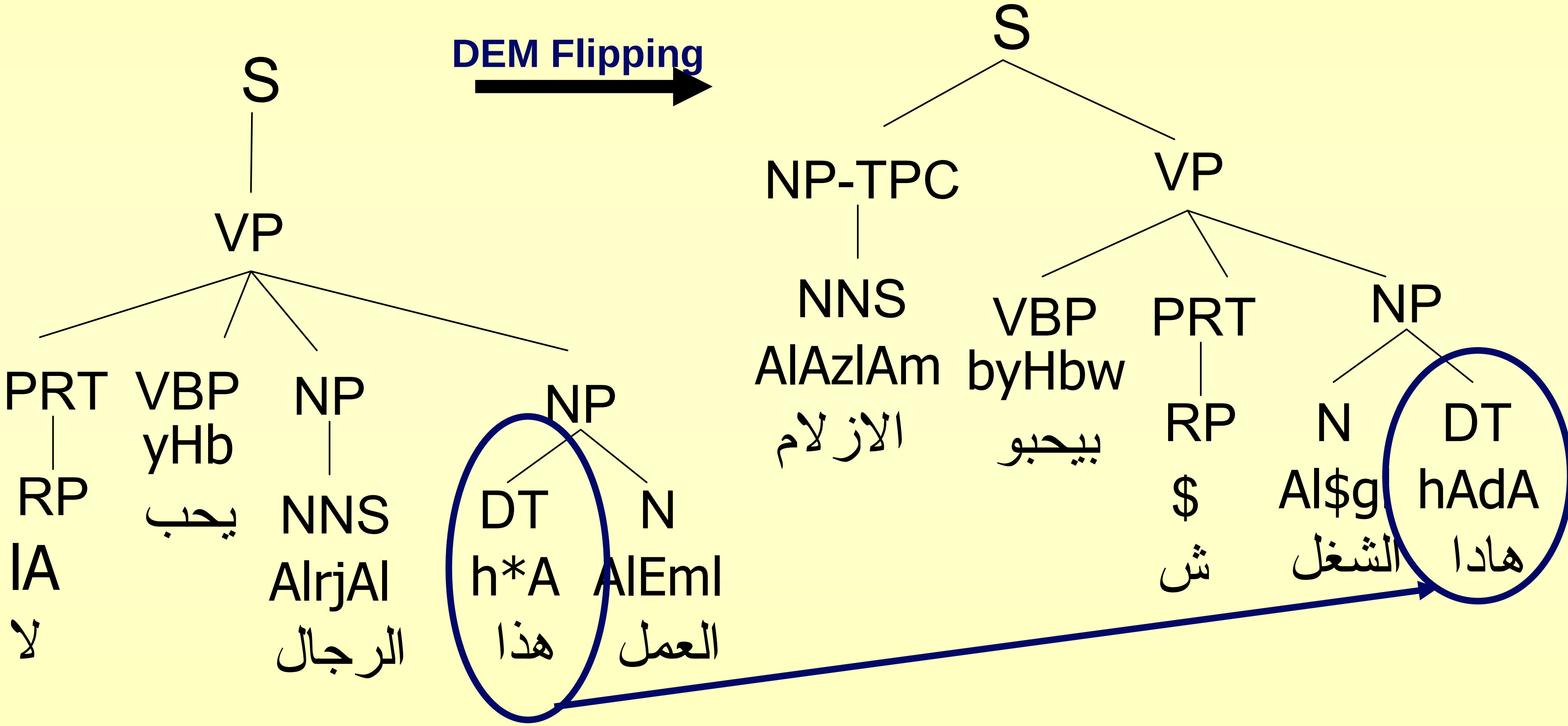
S
VP
PRT VBP NP NP
RP
lA
لا
yHb
يحب
NNS
AlrjAl
الرجال
DT N
h*A AlEml
هذا العمل

S
NP-TPC VP
NNS
AlAzlAm
الازلام
VBP
byHbw
بيحبو
PRT
RP
$
ش
NP
N DT
Al$gl hAdA
الشغل هادا

Syntactic Transformations



Syntactic Transformations



Lexical Transformations

- Using the dictionaries for finding word correspondences from MSA to LEV {Habash}
  - SM: Closed Class dictionary in addition to the 100 most frequent terms and their correspondences
  - LG: SM + open class LEV TB dev set types
- Two types of probabilities associated with entries in dictionary: {Nichols, Sima’an, Hwa}
  - EM probabilities
  - Uniform probabilities

Morphological Manipulations

- Replacing all occurrences of MSA VB ‘want’ to NN ‘bd’ and inserting possessive pronoun
- Replacing MSA VB /lys/ by and RP m$
- Changing VBP verb to VBP b+verb

Experiments

- Tree normalization
- Syntactic transformations
- Lexical transformations
- Morphological transformations
- Interactions between lexical, syntactic and morphological transformations

Parser

- Bikel Parser off-shelf

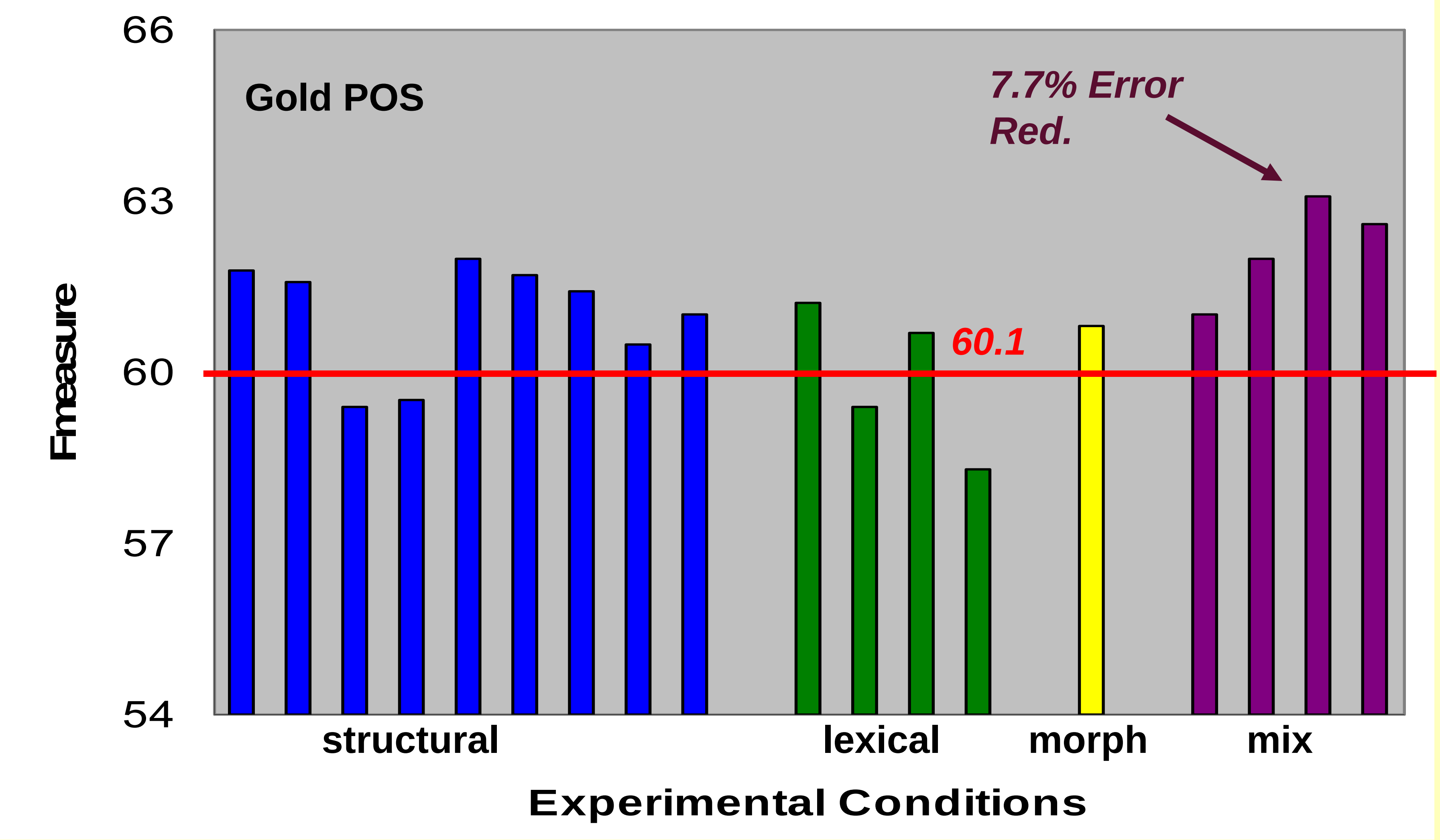
Evaluation

- Labeled precision/Labeled recall/F-measure

Experiment Variations

POS tags	No Lexicon	Small Lexicon	Big Lexicon
None	53.2F		
Automatic			
Gold			

Performance on DevSet



Results

F measure/GoldTag	Dev	Test
Baseline	60.1	60.2
TNORM+NEG	62	61
Lex SM+EMprob	61.2	59.7
MORPH	60.8	60
Lex SM+EMprob +MORPH	61	59.8
TNORM+NEG +MORPH	62	60.6
TNORM+NEG+Lex SM+EM	63.1	61.5
TNORM+NEG+Lex SM+EM +MORPH	62.6	61.2

Observations

- Not all combinations help
- Morphological transformations seem to hurt when used in conjunction with other transformations
- Difference in domain and genre account for uselessness of the large dictionary
- EM probabilities seem to play the role of LEV language model
- Caveat: Lexical resources even for closed class are created for LEV to MSA not the reverse (25% type defficiency in coverage of closed class items)

Conclusions & Future Directions

- Resource consistency is paramount

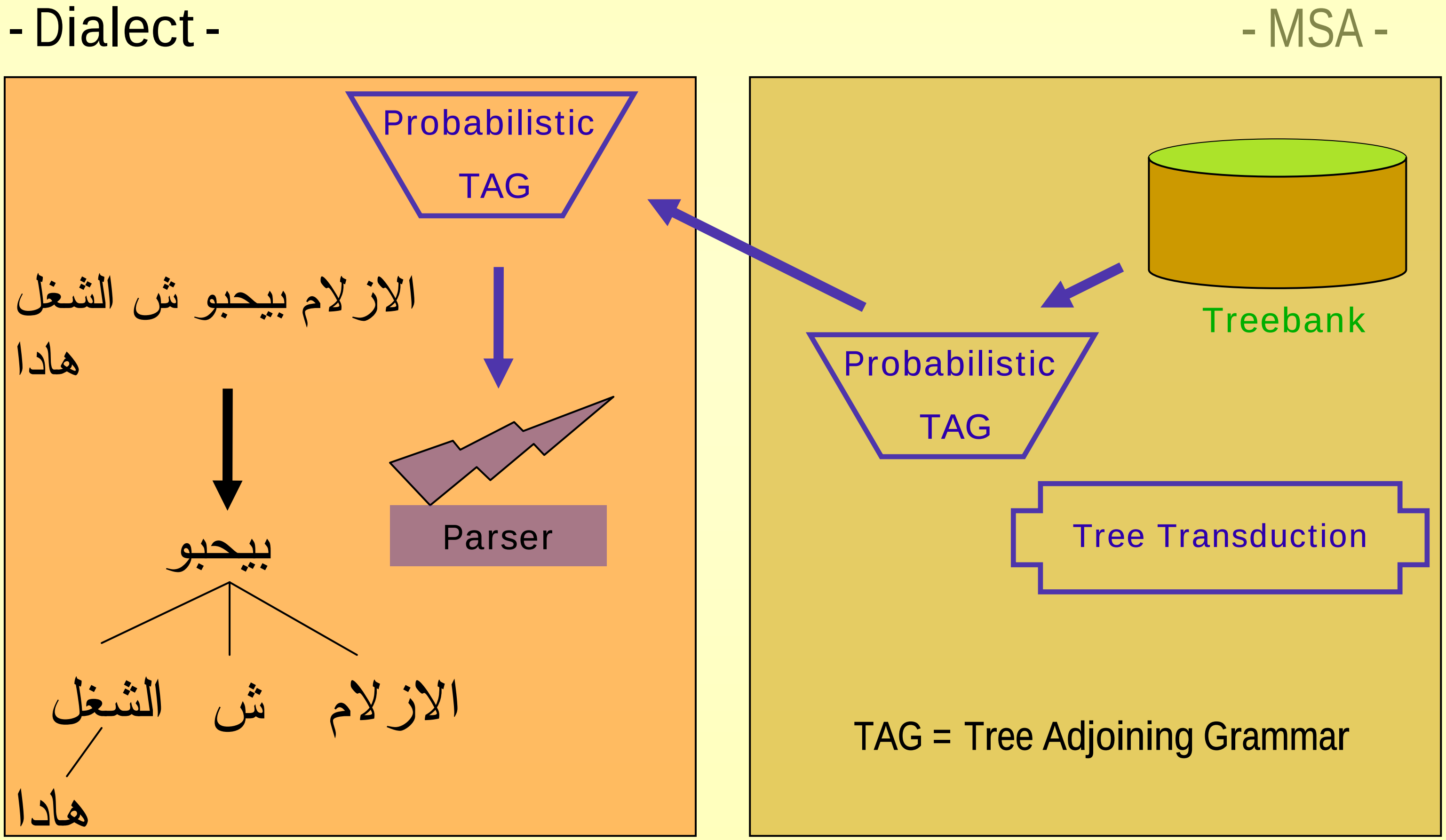
Future Directions

- ❑ More Error analysis
- ❑ Experiment with more transformations
- ❑ Add a dialectal language model
- ❑ Experiment with more balanced lexical resources
- ❑ Test applicability of tools developed here to other Arabic dialects
- ❑ Maybe automatically learn possible syntactic transformations?

Global Overview

- Introduction (Owen Rambow)
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
  - Introduction and Baselines
  - Sentence Transduction
  - Treebank Transduction
  - **Grammar Transduction (David Chiang)**
- Conclusion

Grammar Transduction



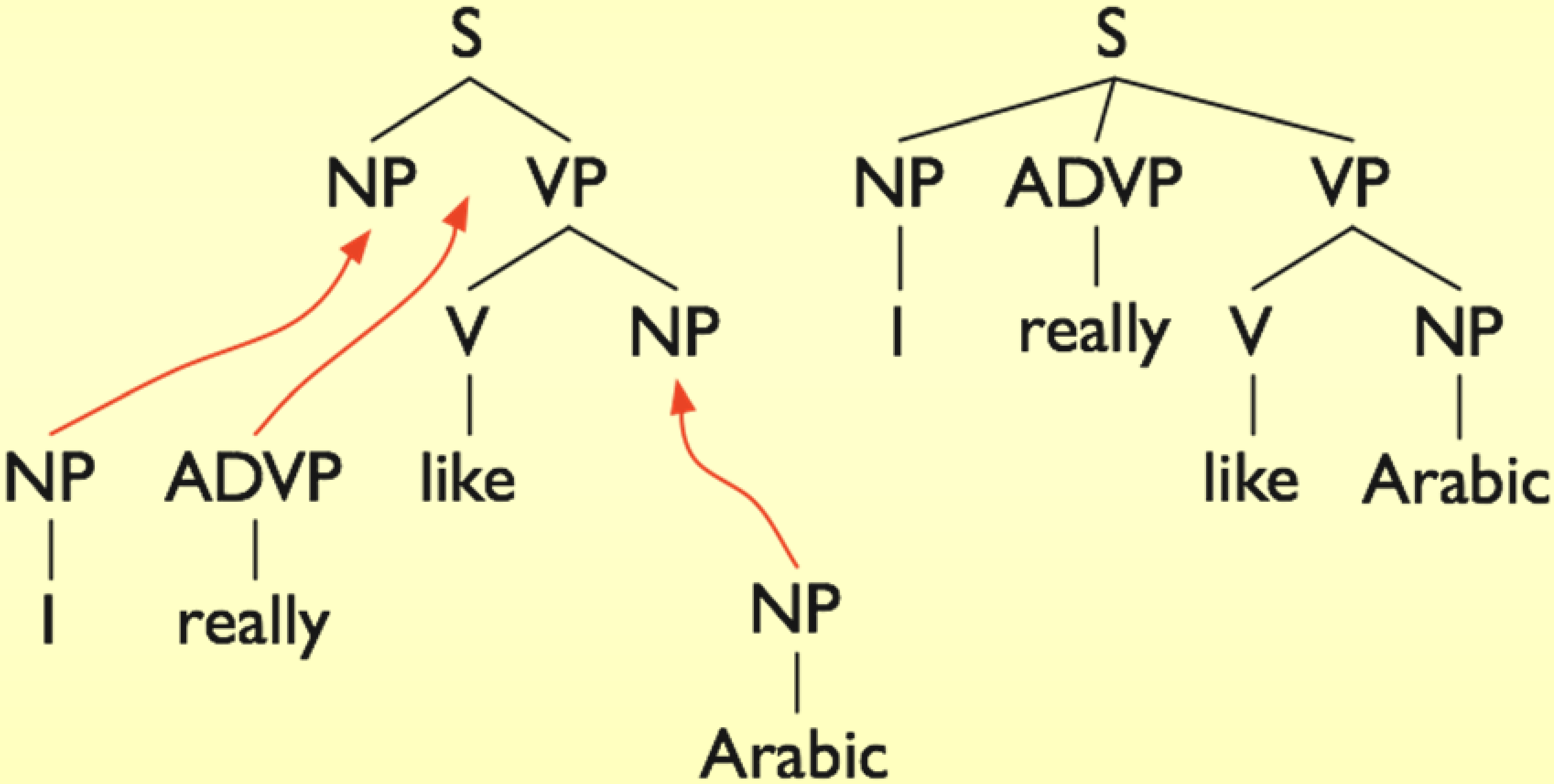
TAG = Tree Adjoining Grammar

Grammar Transduction

- Transform MSA parsing model into dialect parsing model
- More precisely: into an MSA-dialect synchronous parsing model
- Parsing model is defined in terms of *tree-adjoining grammar* derivations

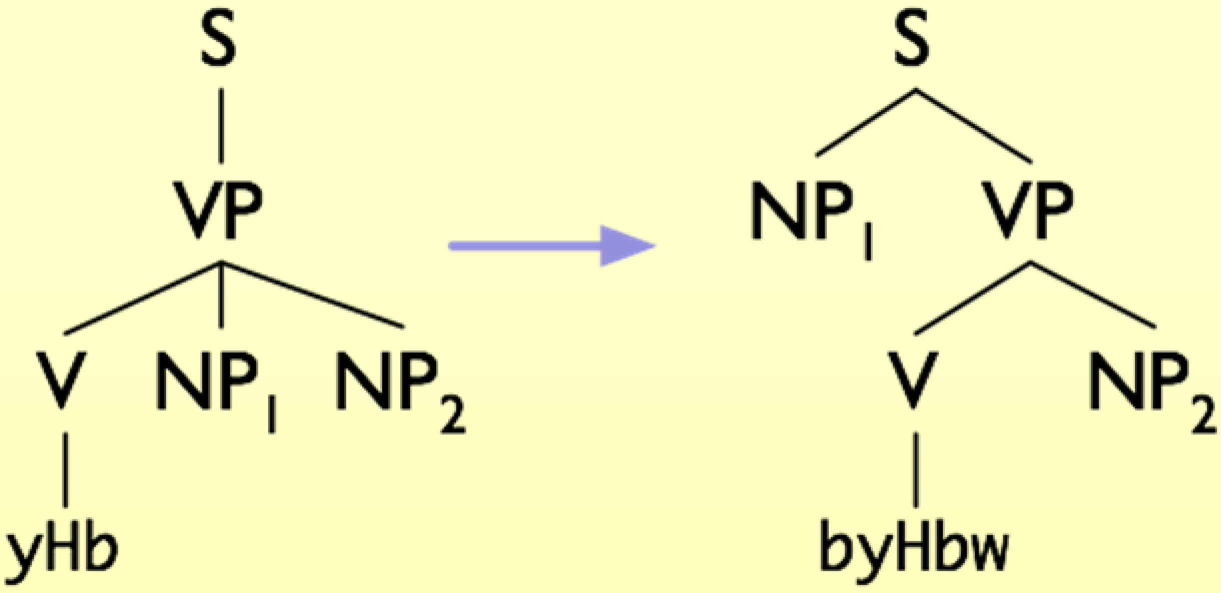
Contributors: David Chiang and Owen Rambow

Tree-Adjoining Grammar



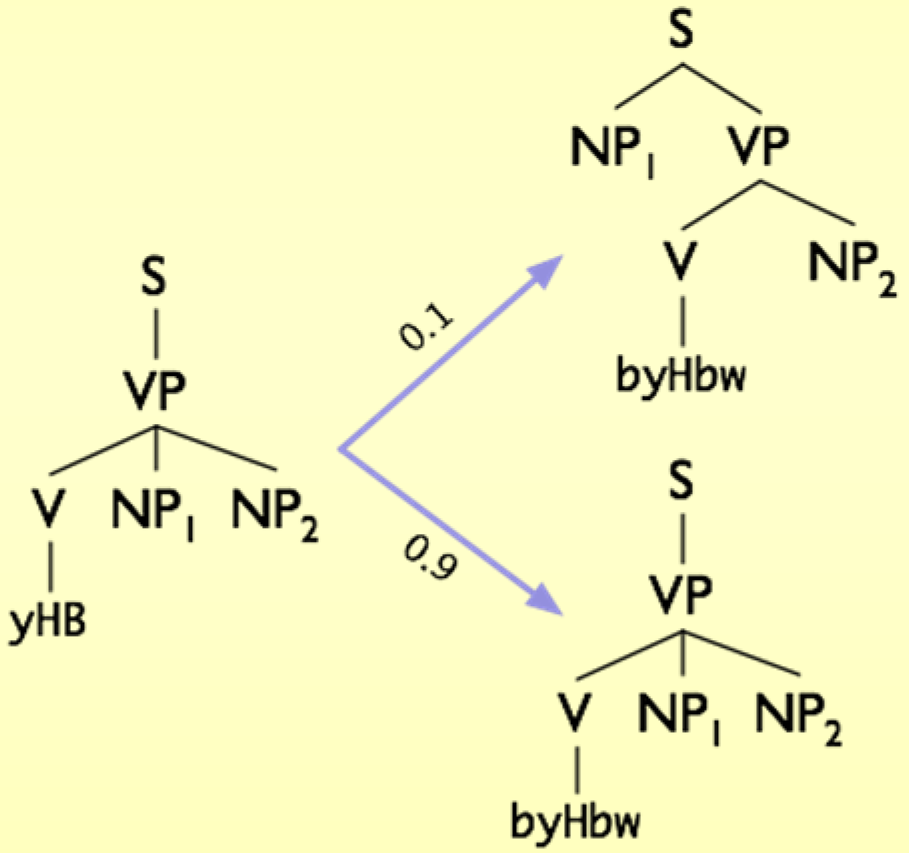
Transforming a TAG

- Thus: to transform a TAG, we specify transformations on elementary trees



Transforming Probabilities

- MSA parsing model is probabilistic, so we need to transform the probabilities too
- Make transformations probabilistic: this gives $P(T_{Lev}|T_{MSA})$



Probability Model

To parse, search for:

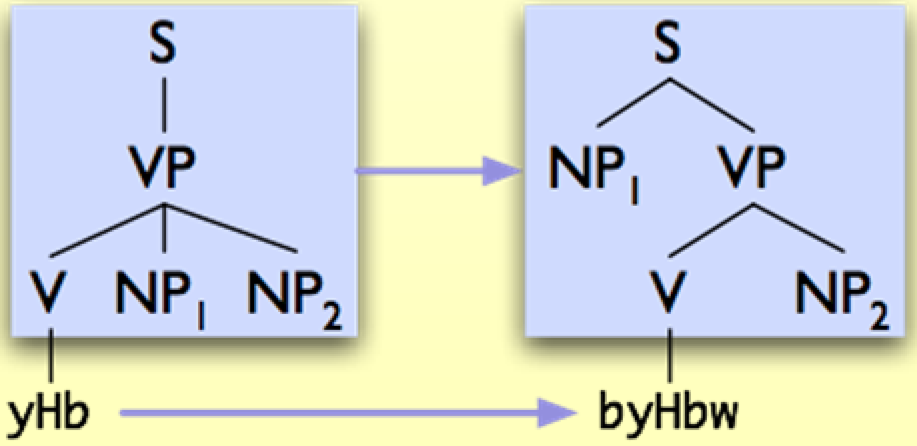
$$\begin{aligned} \arg\max P(T_{Lev}) &\approx \arg\max P(T_{Lev}, T_{MSA}) \\ &= \arg\max P(T_{Lev}|T_{MSA})\, P(T_{MSA}) \end{aligned}$$

given by grammar transformation

learned from MSA treebank

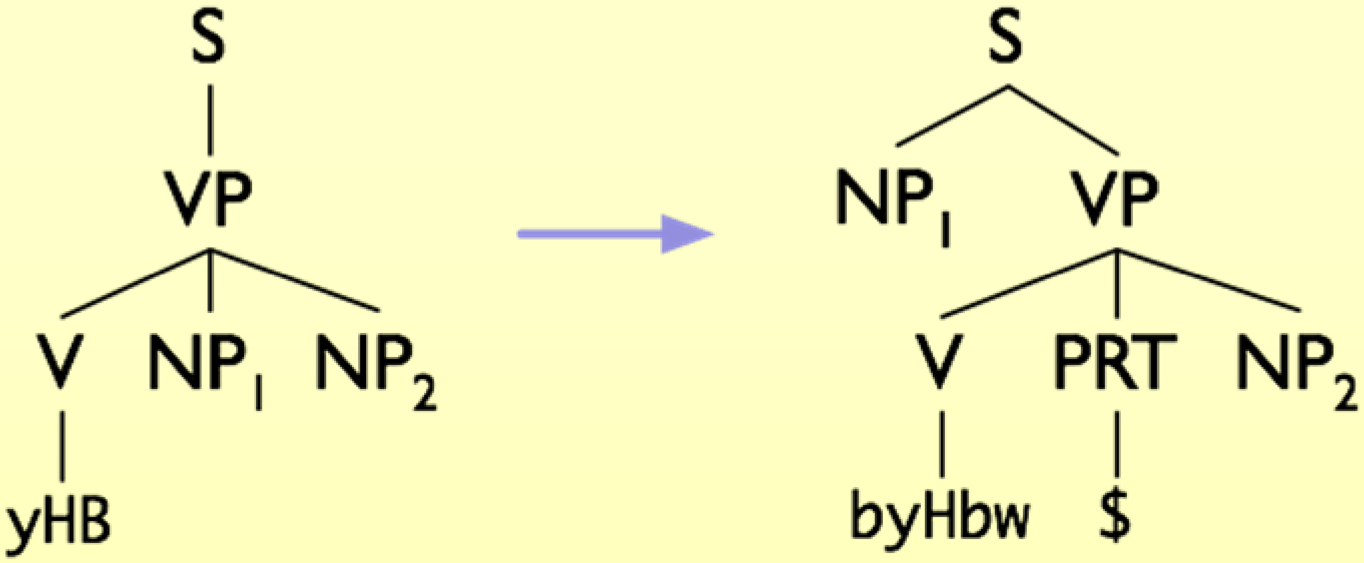
Probability Model

- Full set of mappings is very large, because elementary trees are lexicalized
- Can backoff to translating unlexicalized part and lexical anchor independently



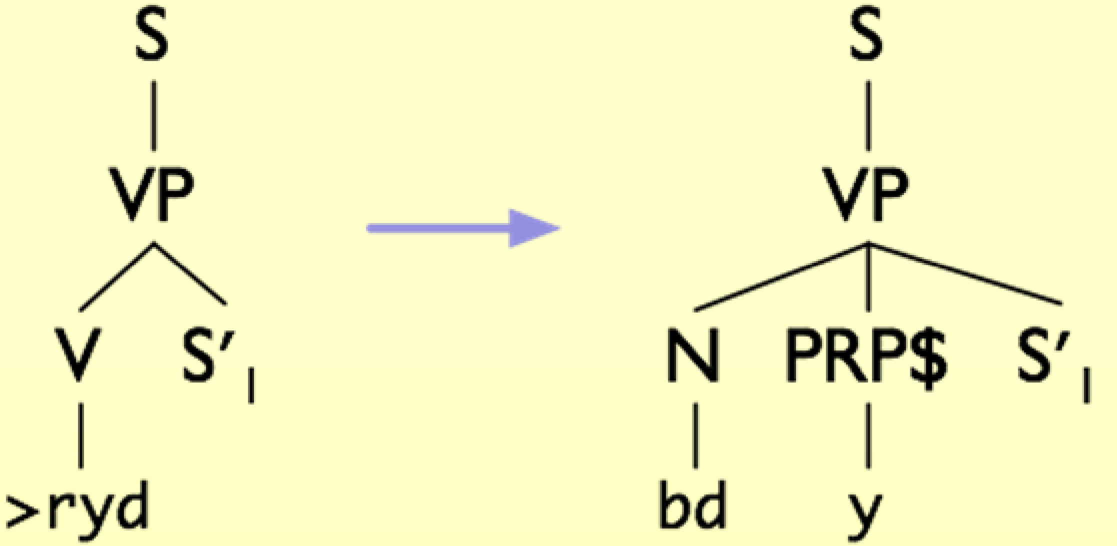
Transformations

- VSO to SVO transformation
- Negation:



Transformations

- ‘want’



Experiments (devtest)

POS tags	No Lexicon	Small Lexicon	Big Lexicon
None			
Automatic			
Gold	□	□	

Results (devtest)

	Recall	Prec	F1
Baseline	62.5	63.9	63.2
Small lexicon	67.0	67.0	67.0
VSO→SVO	66.7	66.9	66.8
negation	67.0	67.0	67.0
‘want’	67.0	67.4	67.2
negation+‘want’	67.1	67.4	67.3

Experiments (test)

POS tags	No Lexicon	Small Lexicon	Big Lexicon
None	□	□	□
Automatic			
Gold			

Results (test)

	Recall	Prec	F1
Baseline	50.9	55.4	53.1
All, no lexical	51.1	55.5	53.2
All, small	58.7	61.8	60.2
All, large	60.0	62.2	61.1

Further Results

- Combining with unsupervised POS tagger hurts (about 2 points)
- Using EM to reestimate either $P(T_{Lev}|T_{MSA})$ or $P(T_{MSA})$
    - no lexicon: helps first iteration (about 1 point), then hurts
    - small lexicon: doesn’t help

Conclusions

- Syntactic transformations help, but not as much as lexical
- Future work:
  - transformations involving multiple words and syntactic context
  - test other parameterizations, backoff schemes

Global Overview

- Introduction
- Student Presentation: Safi Shareef
- Student Presentation: Vincent Lacey
- Lexicon
- Part-of-Speech Tagging
- Parsing
    - Introduction and Baselines
    - Sentence Transduction
    - Treebank Transduction
    - Grammar Transduction
- **Conclusion (Owen Rambow)**

Accomplishments

- Created software for acquiring lexicons from comparable corpora
- Investigated use of different lexicons in Arabic dialect NLP tasks
- Investigated POS tagging for dialects
- Developed three approaches to parsing for dialects, with software and methodologies

Summary: Quantitative Results

- POS tagging
  - No lexicon to small lexicon: 70% to 77%
  - Small lexicon to small lexicon with in-domain information: 77% to 80%
- Parsing
  - No lexicon to small lexicon: 63.2% to 67%
  - Small lexicon to small lexicon with syntax: 67% to 67.3%
  - Train on 10,000 trebanked words: 69.3%

Resources Created

- Lexicons:
  - Hand-created closed-class, open-class lexicons for Levantine
- POS Tagging:
  - Software for adapting MSA tagger to dialect
- Parsing:
  - Sentence-transduction & parsing software
  - Tree-transformation software
  - Synchronous grammar framework
- Treebanks
  - Transduced dialect treebank

Future Work

- Improve reported work
  - Comparable corpora for Arabic dialects
  - Improve POS results
  - Explore more tree transformations for grammar transduction, treebank transduction
  - Include structural information for key words
- Combine leveraging MSA with use of small Levantine treebank
  - Already used in POS tagging
  - Combine transduced treebank with annotated treebank
  - Augment extracted grammar with transformed grammar