

You got your StatMT  
on my rules!

You got your rules  
in my StatMT!

# MoJo

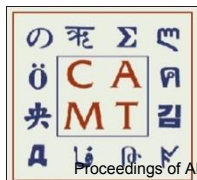
Bringing Hybrid MT to the  
Center for Applied Machine Translation

**Marianna J. Martindale,**  
**Center for Applied Machine Translation (CAMT)**



# MT in Research & Industry vs CAMT

- Research is almost entirely StatMT (now Neural)
  - Emphasis on new techniques
  - Most research on high-resource language pairs (except LORELEI & MATERIAL)
  - Not concerned with operational constraints
- In industry StatMT is the norm (for now)
  - Primarily commercially viable language pairs (high-resource)
  - Speed is important, compute resources may or may not be
- CAMT's GOTS MT is (currently) rule-based
  - Many languages **regardless of resource availability**
  - Speed is important, compute resources limited (server OR laptop)
  - Fidelity is more important than fluency



# Why not StatMT before?

- Technical issues with StatMT
  - Speed
  - Memory
  - Well-engineered systems not readily available
  - Can be tricky to build right



# Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

**OPEN SOURCE**



# Why not StatMT before?

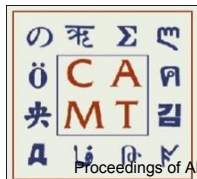
- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

**OPEN SOURCE**

- Domain needs

- Many languages (often low-resource)



# Languages Supported in CyberTrans

AFRIKAANS	DANISH	JAPANESE	PERSIAN	TAJIK
ALBANIAN	<b>DARI</b>	JAVANESE	PERSIAN Romanized	TAJIK Romanized
AMHARIC	DUTCH	KAZAKH	POLISH	<b>TAUSUG</b>
ARABIC	ESTONIAN	KOREAN	PORTUGUESE	<b>TETUM</b>
ARABIC Romanized	FINNISH	KURDISH (Kurmanji)	PUNJABI	THAI
ARMENIAN	FRENCH	<b>KURDISH (Sorani)</b>	ROMANIAN	<b>TOK PISIN</b>
AZERBAIJANI	GALICIAN	KYRGYZ	RUSSIAN	TURKISH
<b>BALUCHI</b>	GEORGIAN	LAO	RUSSIAN Romanized	<b>TURKMEN</b>
BASQUE	GEORGIAN Romanized	LATVIAN	SERBIAN	<b>TWI*</b>
BELARUSIAN	GERMAN	<b>LINGALA</b>	SERBIAN Cyrillic	UKRAINIAN
BULGARIAN	GREEK	LITHUANIAN	<i>SHONA*</i>	UKRAINIAN Romanized
BULGARIAN Romanized	GREEK Romanized	MACEDONIAN	SLOVAK	URDU
CATALAN	HAITIAN CREOLE	MACEDONIAN Romanized	SLOVENE	URDU Romanized
CEBUANO	HAUSA	<b>MAGUINDANAON</b>	SOMALI	<b>UYGHUR</b>
<b>CHAVACANO</b>	HEBREW	MALAGASY	SPANISH	UZBEK Cyrillic
<b>CHECHEN</b>	HINDI	MALAYSIAN	SRANAN	UZBEK Romanized
CHINESE Simplified	HMONG	NORWEGIAN	SUNDANESE	VIETNAMESE
CHINESE Traditional	HUNGARIAN	<b>PAPIAMENTO</b>	SWAHILI	<b>WOLOF</b>
CROATIAN	INDONESIAN	PASHTO	SWEDISH	<b>YAKAN</b>
CZECH	ITALIAN	PASHTO Romanized	TAGALOG	YORUBA



# Why not StatMT before?

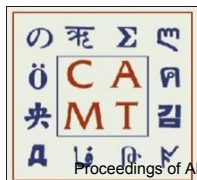
- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

**OPEN SOURCE**

- Domain needs

- Many **languages** (often low-resource)
- Little or no ***in-domain*** parallel text
- Frequent sometimes urgent **updates**
- **Fidelity** as priority (accurate, traceable)



# Why not StatMT before?

- Technical issues with StatMT

- Speed
- Memory
- Well-engineered systems not readily available
- Can be tricky to build right

**OPEN SOURCE**

- Domain needs

- Many **languages** (often low-resource)
- Little or no ***in-domain*** parallel text
- Frequent sometimes urgent **updates**
- **Fidelity** as priority (accurate, traceable)

## MoTrans

- Human instead of bitext
- Updated based on actual text submitted
- Easy to trace input to output
- Caveat: Sacrifice fluency for fidelity





# Features of Rule-based and StatMT

## Rule-based

- Rules are composed by language experts
- Performs a deep source language analysis
- Easy to update, adapt to new domains
- Easy to trace input to output
- Very fast

## StatMT

- Learns automatically from example translations
- Doesn't require language-specific knowledge
- Leverages Big Data
- More fluent translations
- Recent engineering advances make adoption easier

Best of both worlds?



# Best of both worlds

## MoTrans

Human constructed  
 Domain focused  
 Knowledge-rich  
 CAMT linguistic and  
 technical investment

## Statistical

Learned automatically  
 Generic  
 Language-agnostic  
 Commercially dominant  
 Open source

## Hybrid



# Example (Russian)

System	Output
Motrans	He noted, that presidential pre-election campaign provoked “discrepant and often frequently vulgar rhetoric,” eating away democracy and society.
StatMT	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” разъедающую democracy and society .
Hybrid	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” eating away democracy and society.
Human	He said the presidential campaign has brought “divisive and often vulgar rhetoric” that corrodes democracy and society.



# Example (Russian)

System	Output
Motrans	From Moscow to Sochi on the train about two days! Really? Is it possible? You want to lead two days in the uncomfortable train?
StatMT	From Moscow to Sochi to train about two days! Do you want to spend two days in awkward train?
Hybrid	From Moscow to Sochi on the train about two days! Do you want to spend two days in uncomfortable train?
Human	From Moscow to Sochi by train is close to 2 days! Do you really want to spend two days in an uncomfortable train?



# Example (Swahili)

System	Output
Motrans	LABLA America in/at what region? America is big.
StatMT	“Maybe America in what state? The United States is the greatest.
Hybrid	Maybe America in what region? The United States is big.
Human	To be more precise, which state in America? America is vast.

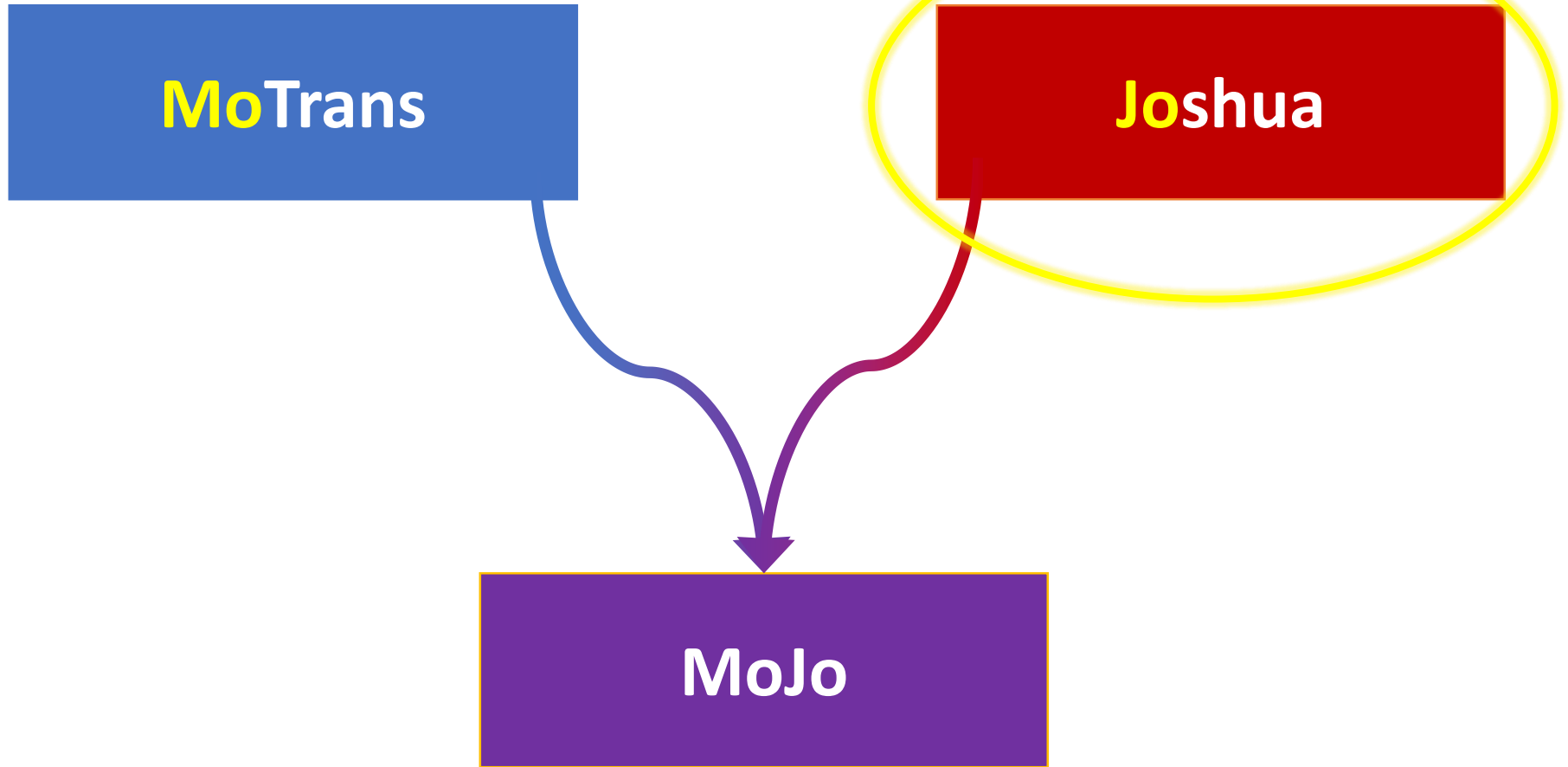


# Example (Swahili)

System	Output
Motrans	Pat: say (you/it) succeeded to get children and MUMEO Mohammed HAMIS?
StatMT	Pat: is ulifanikiwa to children and mumeo Mohamed Hamis?
Hybrid	Pat: <i>Have you</i> succeeded in <i>getting</i> children mumeo Mohamed Hamis?
Human	Pat: Were you successful at having children by that husband Mohamed Hamis?



# Best of both worlds

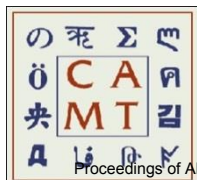


# Apache Joshua



- Open-source Java statistical machine translation system
- Apache project currently in Incubation Stage
- Provides both phrase-based and hiero StatMT
- Pre-built language packs available for download
- **Written in Java** (like CyberTrans)
- **Development lead conveniently located at JHU HLT COE in nearby Baltimore**

<http://joshua.apache.org/>





Best of both worlds

MoTrans

Joshua

MoJo



# MoTrans Translator

Settings	Lexical Entries	Grammar
- , ; / . ' ( @ \$ \u20AC % + < =   0 1 2 3 4 5 6 7 8 9 A Á À Ã Ä Å		
+ - × ÷ ↵ ↶ ↷		
Source *	POS	Target
gabarit^ de fraisage	N	milling jig^
gabarit^ de membrure	N	frame mould^
gabarit^ de mécanicien	N	engineer's jig^
gabarit^ de montage	N	assembly jig^
gabarit^ de perçage	N	drill template^
gabarit^ de traçage	N	contour template^
gabarit^ de vérification	N	inspection gauge^
gabarit-obstacle	N	minimum dimensions
gabarre	N	lighter
gabbro	N	gabbro
gabegie	N	intrigue
gabelage	N	time^ during which the salt was in store t
gabelleur	N	customs official^
gabelier	N	salt-tax officer^
gabelle	N	salt tax^
gaber	V	joke



- Morphological Translator
- Fast
- Deep morphological analysis
- Expressive lexicon and grammar
- Continually updated by lexicographers
- Quick “better than nothing” for Low Resource Languages
- Currently over 40 languages
- Many users, positive feedback

# MoTrans Lexicon Example

- Lexical entries

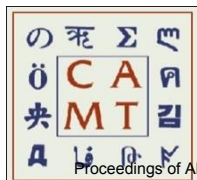
- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural



# MoTrans Lexicon Example

- Lexical entries

- **puntilla** |N| lace
- **pas/ar** |V.AR| pass
- **de** |DE| of
- **pas de puntillas** |V.AR| sidestep |S--

## Source

Usually lemma

But not always

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural



# MoTrans Lexicon Example

- Lexical entries

- puntilla | **N** | lace
- pas/ar | **V.AR** | pass
- de | **DE** | of
- pas de puntillas | **V.AR** | sidestep | S--

Part of Speech

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural



# MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR| sidestep|S--

Target

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural



# MoTrans Lexicon Example

- Lexical entries

- puntilla | N | lace
- pas/ar | V.AR | pass
- de | DE | of
- pas de puntillas | V.AR | sidestep | S--

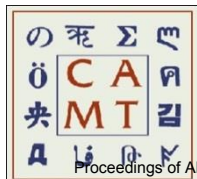
Stem indicator

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural



# MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR| sidestep|S--

## Affix Location Pattern

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural





# MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR|sidestep|S--

- Rules

**PROCLITIC**:V.CONJ:ha:has %s

**SUFFIX**:V.AR:ado:%s:+pastp:V.CONJ

**SUFFIX\_PATTERN**:N:(\*V)(X-θ{s}):%s:+plural

Type



# MoTrans Lexicon Example

- Lexical entries

- puntilla|N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas|V.AR|sidestep|S--

- Rules

PROCLITIC:**V.CONJ**:ha:has %s

SUFFIX:**V.AR**:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:**N**:(\*V)(X-θ{s}):%s:+plural

Part-of-Speech



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ: **ha**:has %s

SUFFIX:V.AR: **ado**:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N: **(\*V)(X-θ{s})**:%s:+plural

Source  
Transformation



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s  
 SUFFIX:V.AR:ado:%s:+pastp:V.CONJ  
 SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:+plural

Target  
Transformation



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

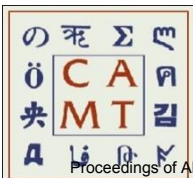
- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:**+pastp**:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:**+plural**

Target  
Conjugation



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:**V.CONJ**

SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:+plural

New  
Part-of-Speech



# MoTrans Lexicon Example

- Lexical entries

- **puntilla** |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

PROCLITIC:V.CONJ:ha:has %s

SUFFIX:V.AR:ado:%s:+pastp:V.CONJ

SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:**+plural**

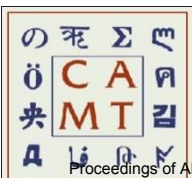
For a noun  
that ends  
in a vowel

Add an 's'

Pluralize  
the English

puntillas ->

lace



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- **pas/ar** |V.AR| **pass**
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--

- Rules

**PROCLITIC:V.CONJ:ha:has %s**

**SUFFIX:V.AR:ado:%s:+pastp:V.CONJ**

**SUFFIX\_PATTERN:N:(\*V)(X-0{s}):%s:+plural**

puntillas ->

lace

ha pasado ->

has passed





# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- **pas de puntillas |V.AR| sidestep |S--**

- Rules

**PROCLITIC:V.CONJ:ha:has %s**

**SUFFIX:V.AR:ado:%s:+pastp:V.CONJ**

**SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:+plural**

puntillas ->

lace

ha pasado ->

has passed

ha pasado de puntillas ->

has sidestepped



# MoTrans Lexicon Example

- Lexical entries

- puntilla |N| lace
- pas/ar |V.AR| pass
- de |DE| of
- pas de puntillas |V.AR| sidestep |S--**

- Rules

**PROCLITIC:V.CONJ:ha:has %s**

**SUFFIX:V.AR:ado:%s:+pastp:V.CONJ**

**SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:+plural**

**puntillas ->**

PUNTILLA

(SUFFIX\_PATTERN:N:(\*V)(X-θ{s}):%s:+plural):N: **lace**

**ha pasado ->**

PAS (PROCLITIC:V.CONJ:HA:has %s:+pastp)

(SUFFIX:V.AR:ADO:%s:+pastp):

V.CONJ: **has passed**

**ha pasado de puntillas ->**

PAS DE PUNTILLAS

(PROCLITIC:V.CONJ:HA:has %s:+pastp)

(SUFFIX:V.AR:ADO:%s:+pastp):

V.CONJ: **has sidestepped**



# Best of both worlds

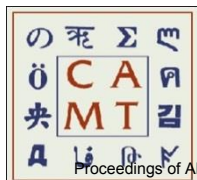
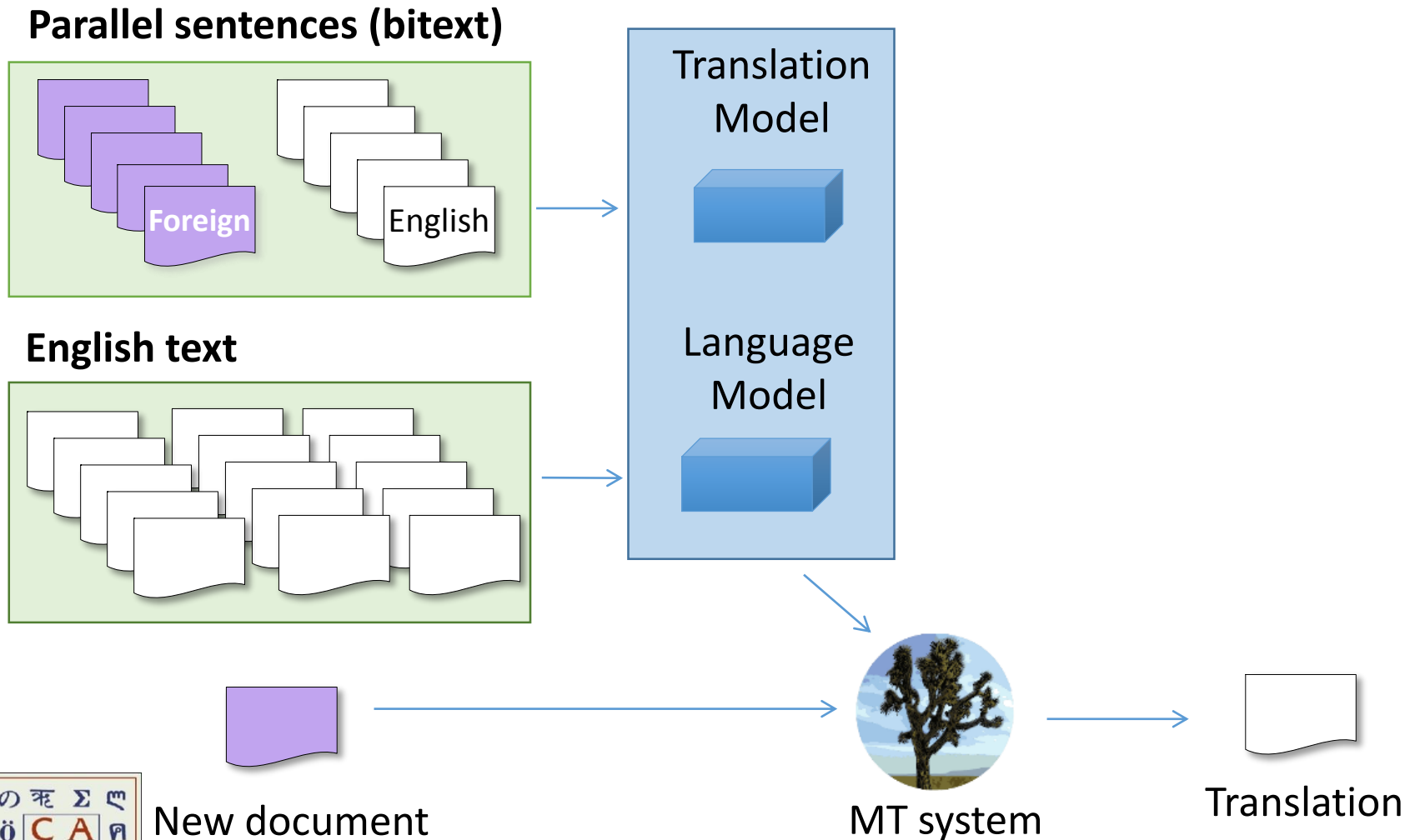
**MoTrans**

**Joshua**

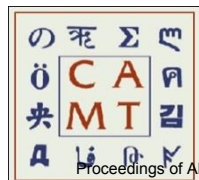
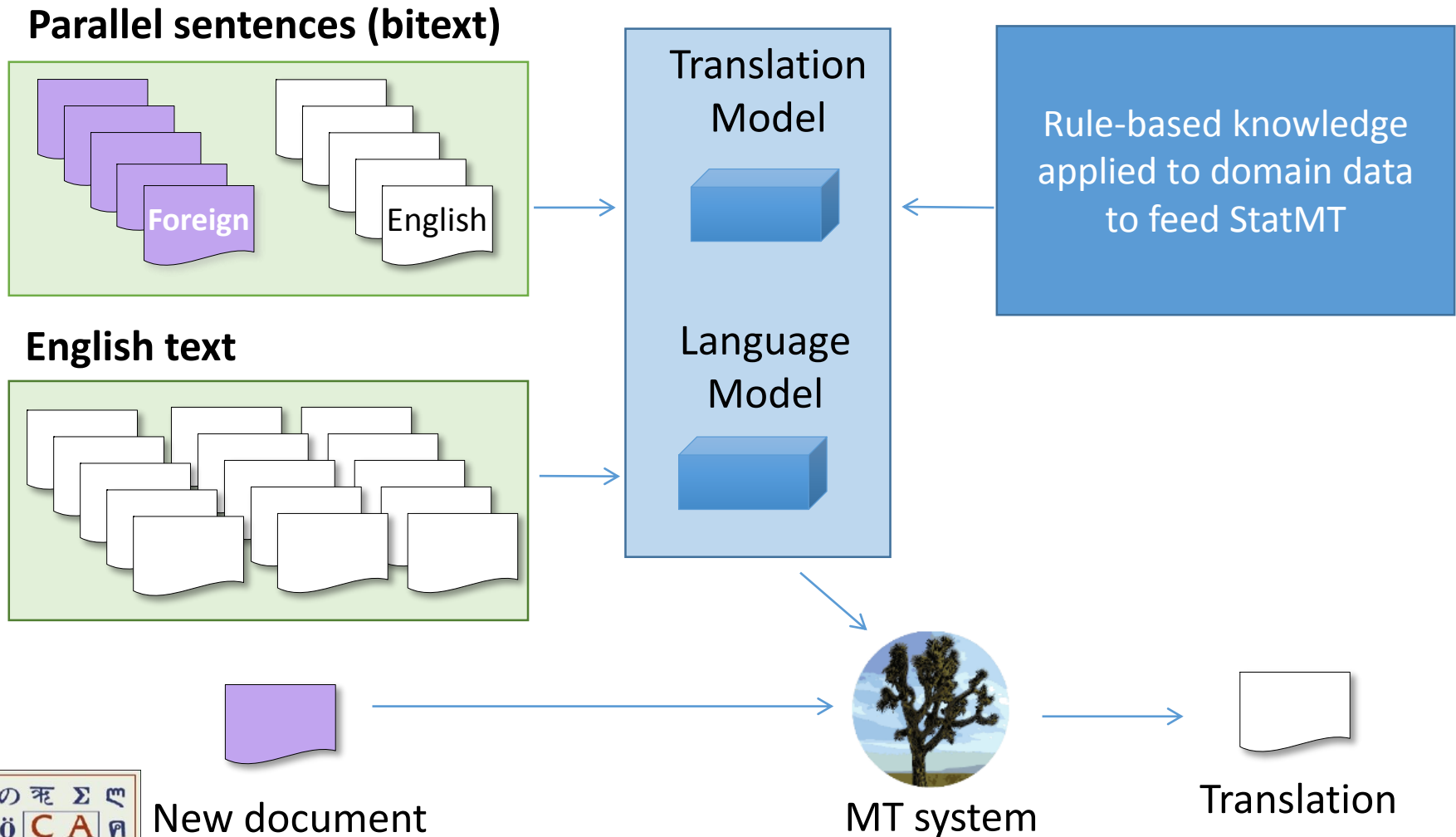
**MoJo**



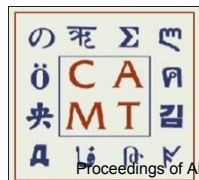
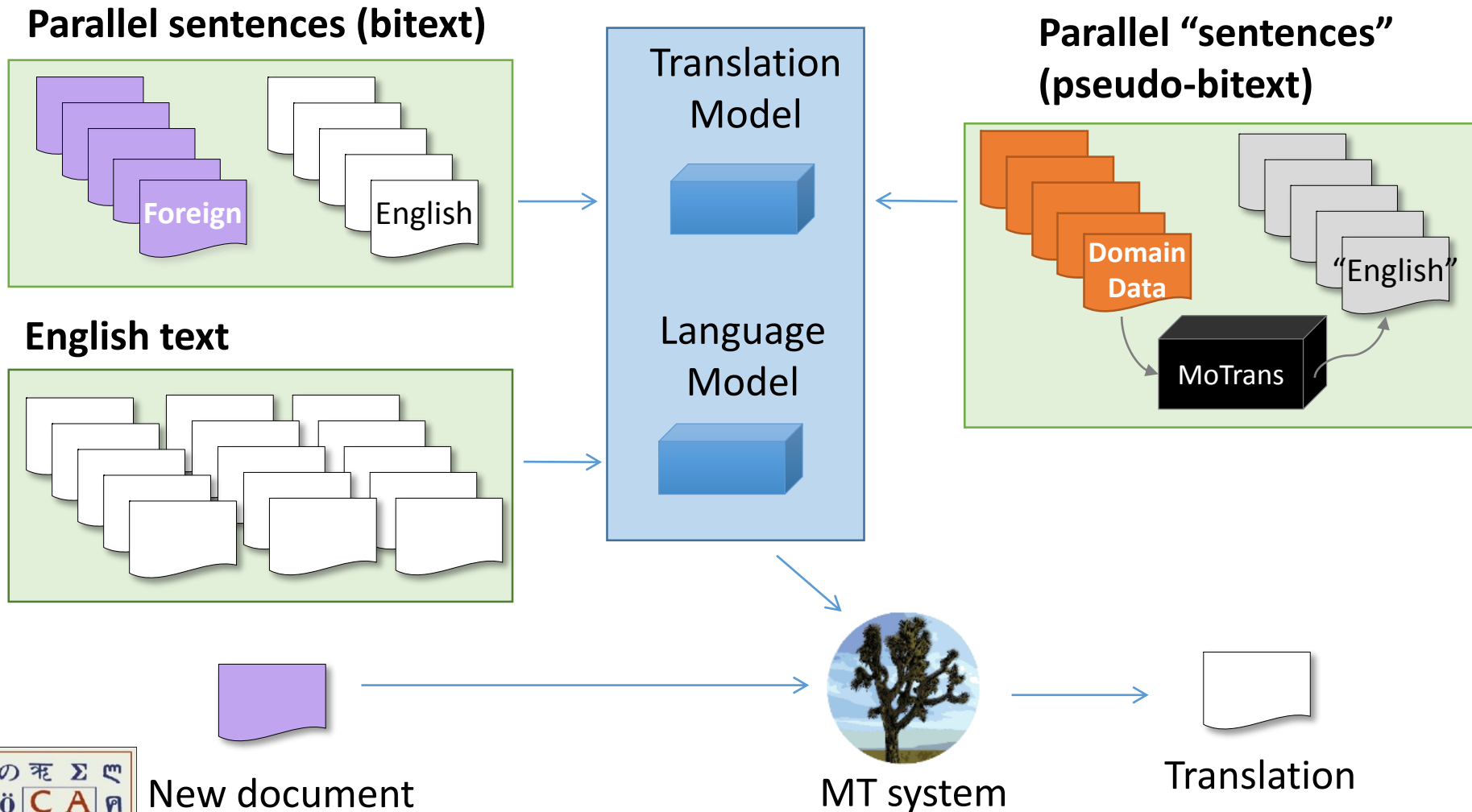
# Building the Hybrid: Base StatMT



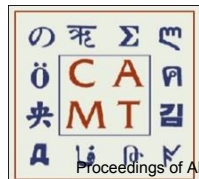
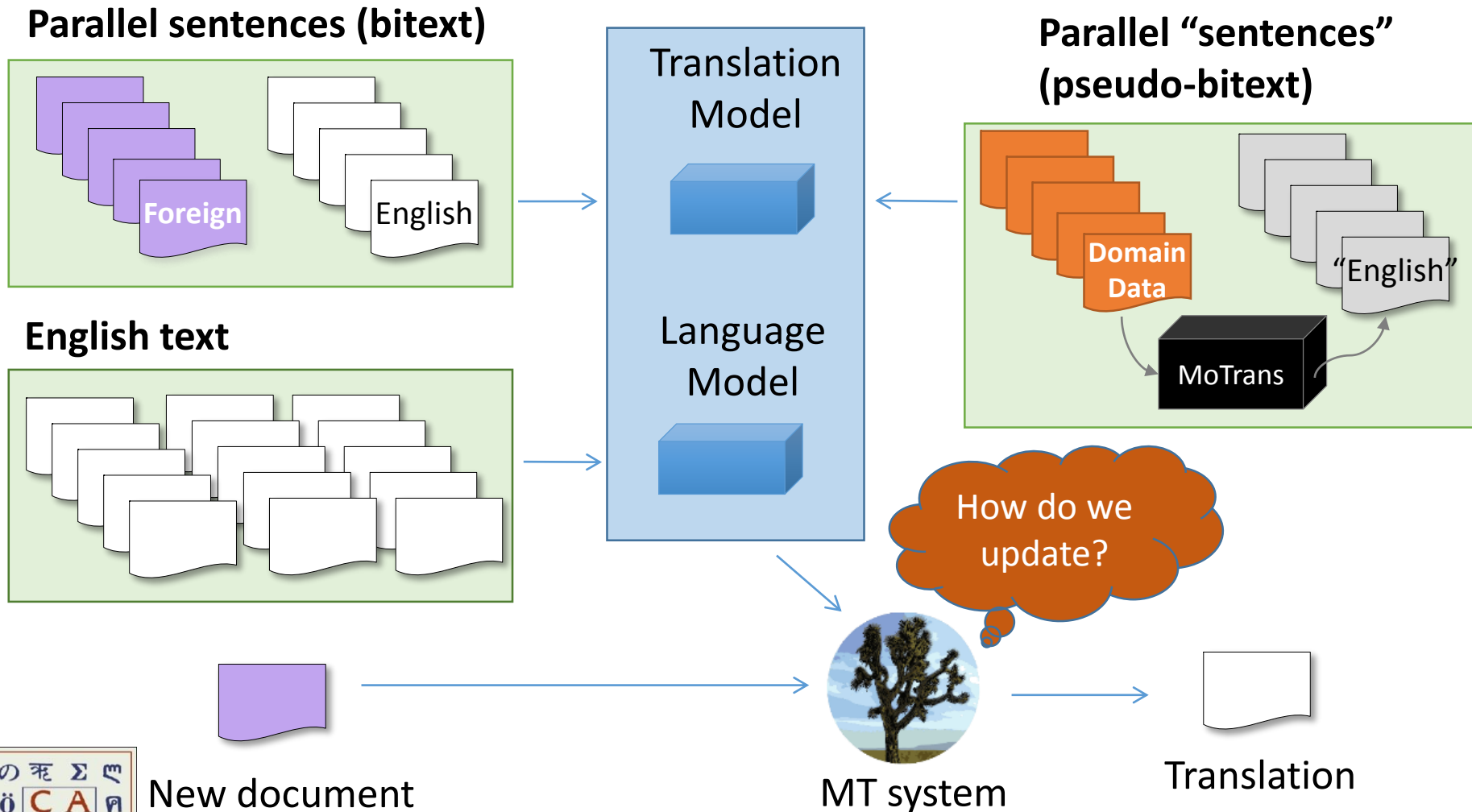
# Building the Hybrid: Black Box



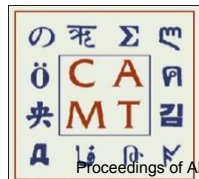
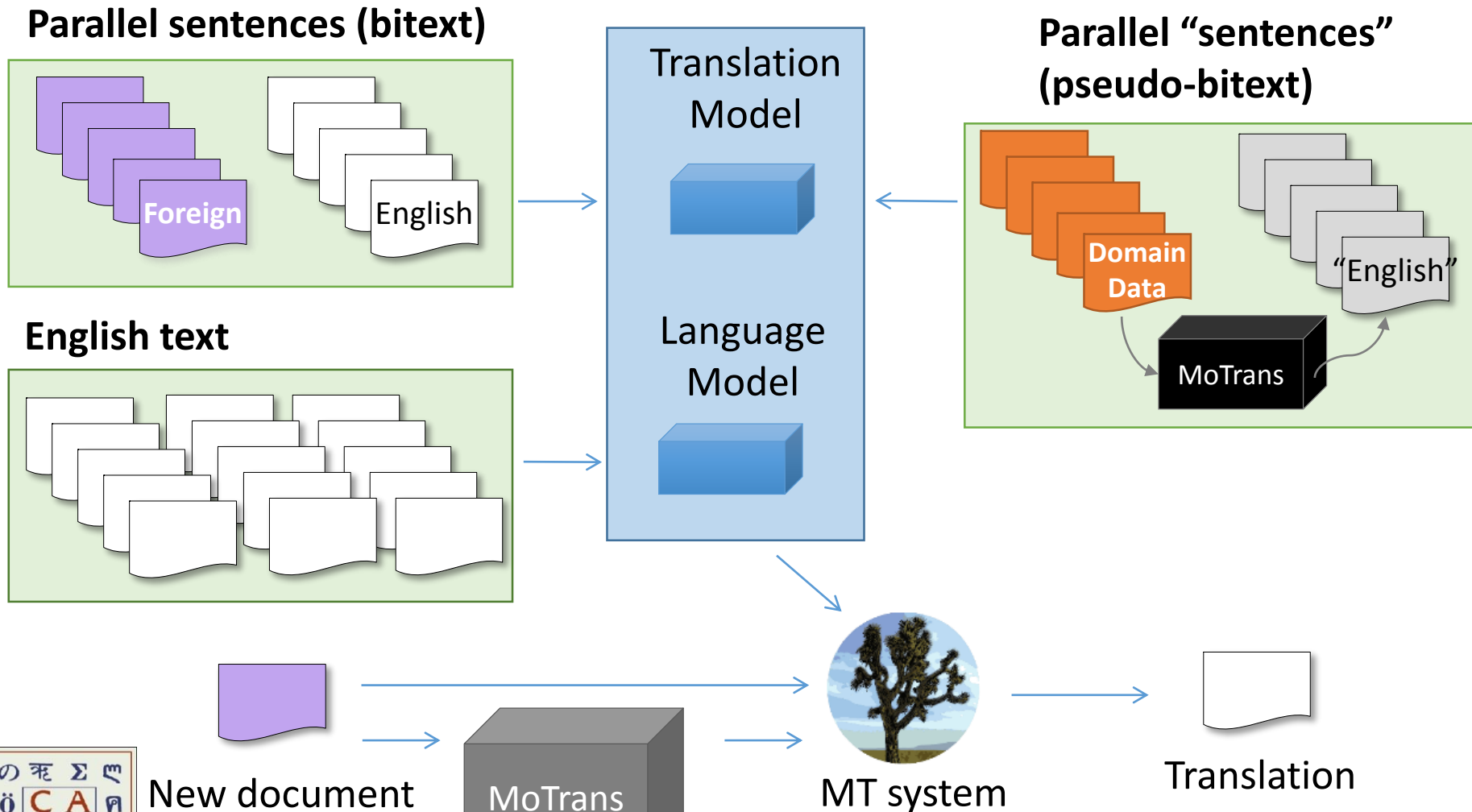
# Building the Hybrid: Black Box



# Building the Hybrid: Black Box

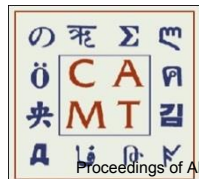
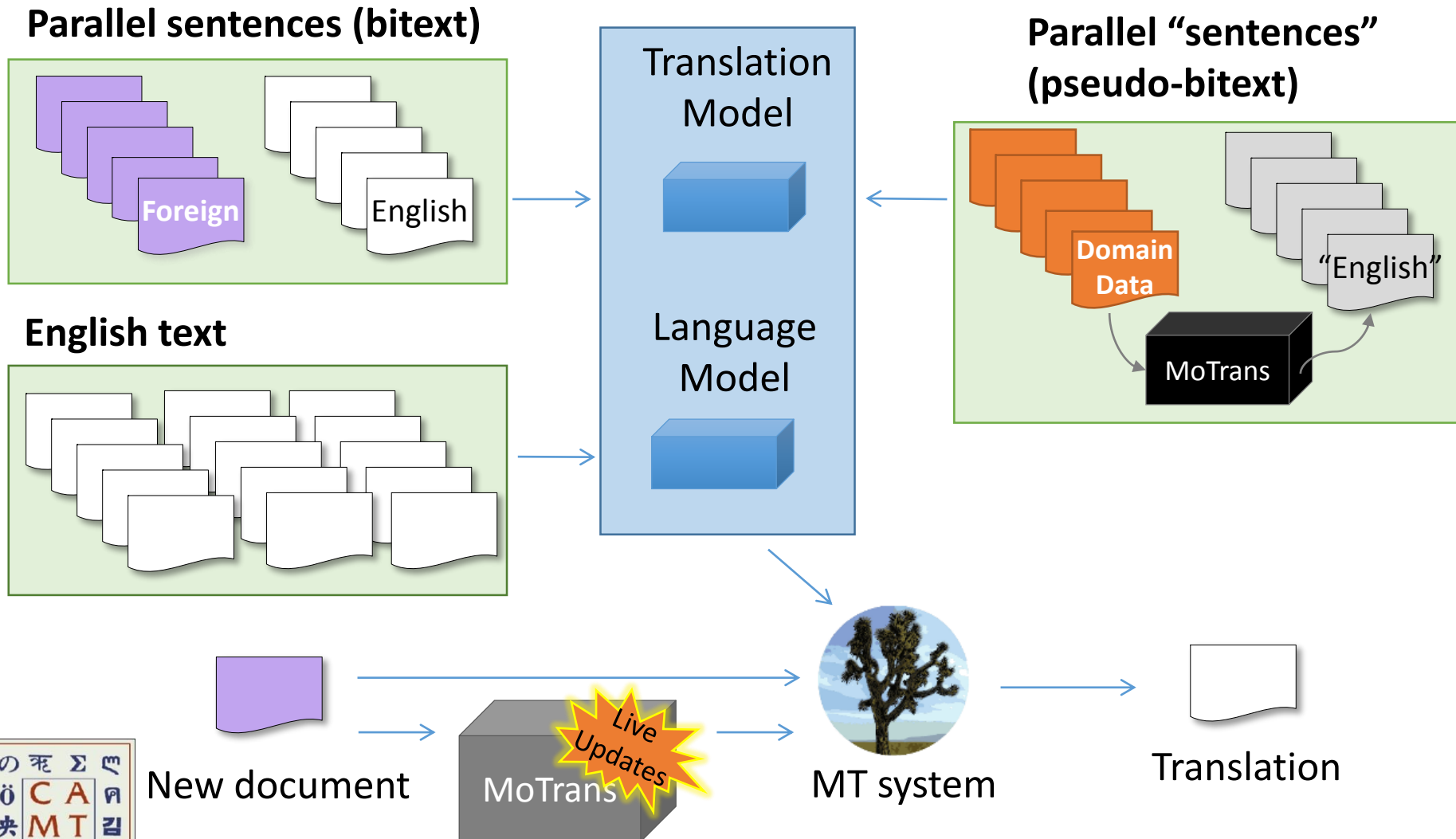


# Building the Hybrid: On Demand

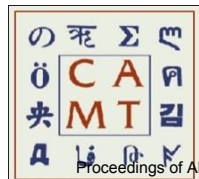
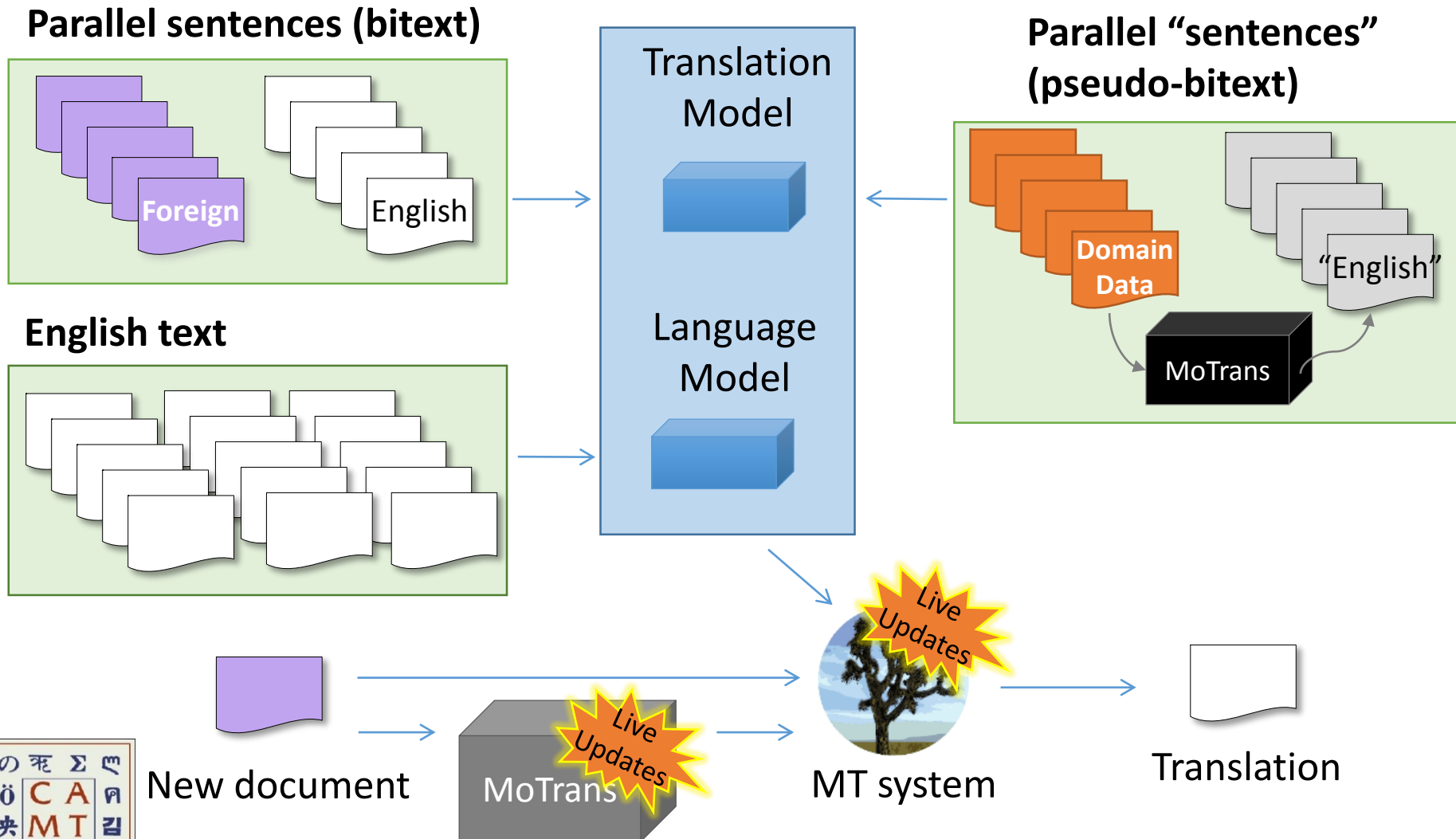




# Building the Hybrid: Live Updates



# Building the Hybrid: Live Updates



# User View

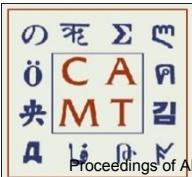
**Translation Options**

- Reset Add 2nd Pass User Settings
- Input Language: Spanish
- Output Language: English
- Encoding: Unknown
- Translator: Recommended (D)
- Topic Dictionary: n/a
- User Dictionary: n/a
- Output Format: Auto Select
- Tab Window File
- Annotations:
- Multi-Pass Options: (0)

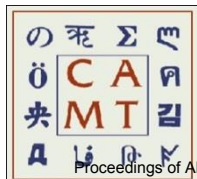
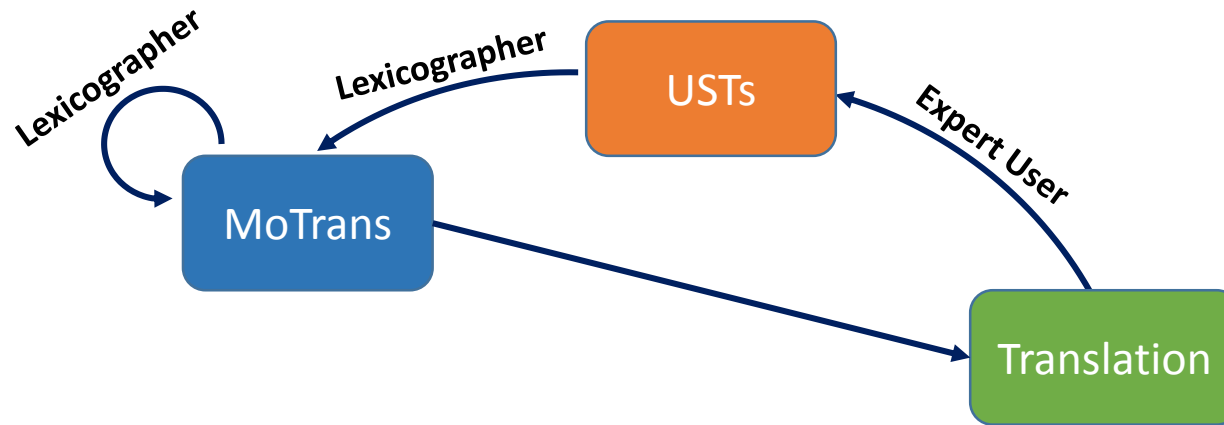
**Text to Translate Translate-1**

Original Text	Motrans	MoJo
<p>En el primer debate de los precandidatos republicanos a la presidencia de los Estados Unidos organizado y transmitido por la cadena estadounidense FOX News el 7 de diciembre de 2015, el Precandidato mantuvo una postura que fue objeto de una gran polémica. Destacó que el sistema político de su país se encuentra presuntamente «roto» y que él y los Estados Unidos «no tienen tiempo de ser políticamente correctos», argumento sustentado en que el país ha perdido protagonismo y competitividad en el escenario global, esto según declaraciones del propio magnate. Él también destacó que no descartaría la posibilidad de ser un candidato independiente a la presidencia del referendo país si no llegase a ser nominado formalmente como "Candidato Presidencial por el Partido Republicano" lo cual fue objeto de críticas dentro de las filas de dicho partido.</p>	<p>In the first debate of the republican primary candidates to the presidency of the organized United States and transmitted by the American chain FOX News the 7 December 2015, the maintained primary candidate a posture that was object of a great controversy. Emphasized that the political system of their country is presumably «I rotate» and that he and the United States «they haven't time of to be politically correct», argument sustained in which El País has lost prominence and competitiveness in the global stage, this according to statements of the own magnate. He also emphasized that wouldn't reject the possibility of to be an independent candidate to the presidency of the referenced country or else came to be nominated formally as "Presidential Candidate by/for the Republican party" which was I object of criticism within the strings of said left.</p>	<p>In the first debate of the republican pre-candidates to the presidency of the United States organized and transmitted by the American chain Fox News the 7 of December of 2015, the Pre-candidate maintained a position that was object of a great controversy. It emphasized that the political system of its country is presumably «broken», and that it and the United States «do not have time to be politically correct», argument sustained in which the country is lost protagonism and competitiveness in the global scene, this according to declarations of the own tycoon. It also emphasized that he would not discard the possibility of being an independent candidate to the presidency of the referred country if he did not arrive to be name formally as "Presidential Candidate by the Republican Party", which was object of critics within the rows of this party.</p>

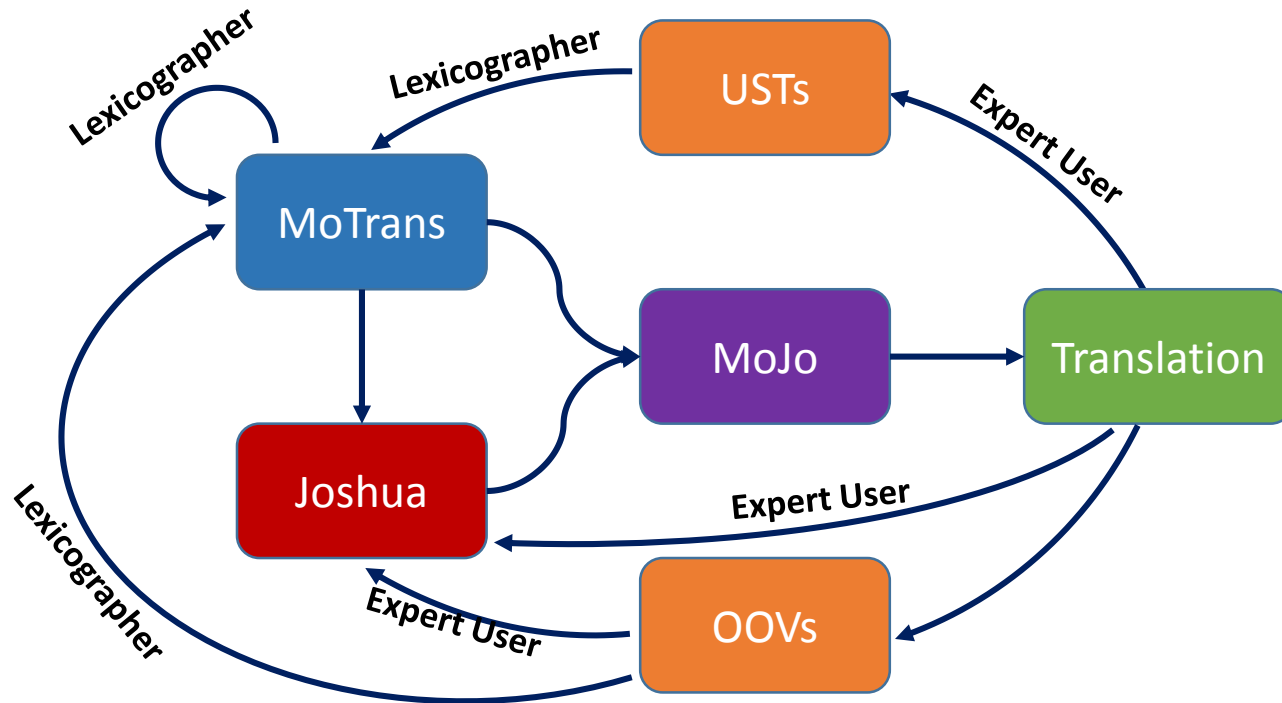
你好 hello 翻译 fan yi windows RQE



# Workflow



# Workflow



# Conclusions

- Mojo online in CyberTrans “soon”
  - Productization in progress
    - Starting with Spanish
    - Other languages will follow
  - Shortly thereafter will be available as add-on for CyberTrans distributions

## Questions?



# Backup Slides



# Example (Portuguese)

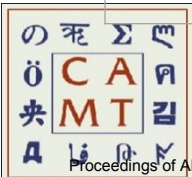
System	Output
Motrans	however, the balloting already finished but not yet there are results ends.
StatMT	Meanwhile, the ballot finished but there is still no final results.
Hybrid	However, the ballot has finished but there is still no final results.
Human	However, the audit is over but there still are no final results.





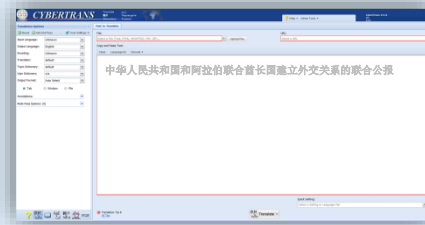
# Example (Arabic)

System	Output
Motrans	And adds Dr. Syrian/Suri that this is reason viewing it to some the patients/al-Murdi who come to him after the passing of the time of their treatment, in addition to the resistance of their bodies to drugs specific/Mu'inah.
StatMT	He said d. Sorry, this is a reason to see some patients who come to him after it was too late for treatment, as well as to resist their bodies certain drugs.
Hybrid	D. Syrian adds that this is the reason see some patients who come to him after the passing of time their treatment, as well as to resist their bodies specific drugs.
Human	Dr. Sory also stated that is why he sometimes had diseased people come in when it was too late for treatment and why there was resistance to certain drugs.



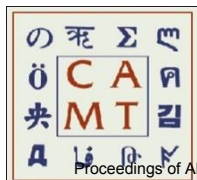
# Center for Applied Machine Translation

中华人民共和国和阿拉伯联合酋长国建立外交关系的联合公报



The Peoples Republic of China and the United Arab Emirates established diplomatic relations with the joint communique

- DoD-recognized Center of Excellence for Machine Translation
  - Serving the US Government for over 14 years.
- Flagship product: **CYBERTRANS**
  - Integrated suite of automated tools for MT, language and encoding identification, spelling and text enhancement, and encoding conversion.



# CyberTrans Usage

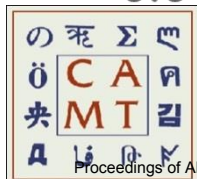
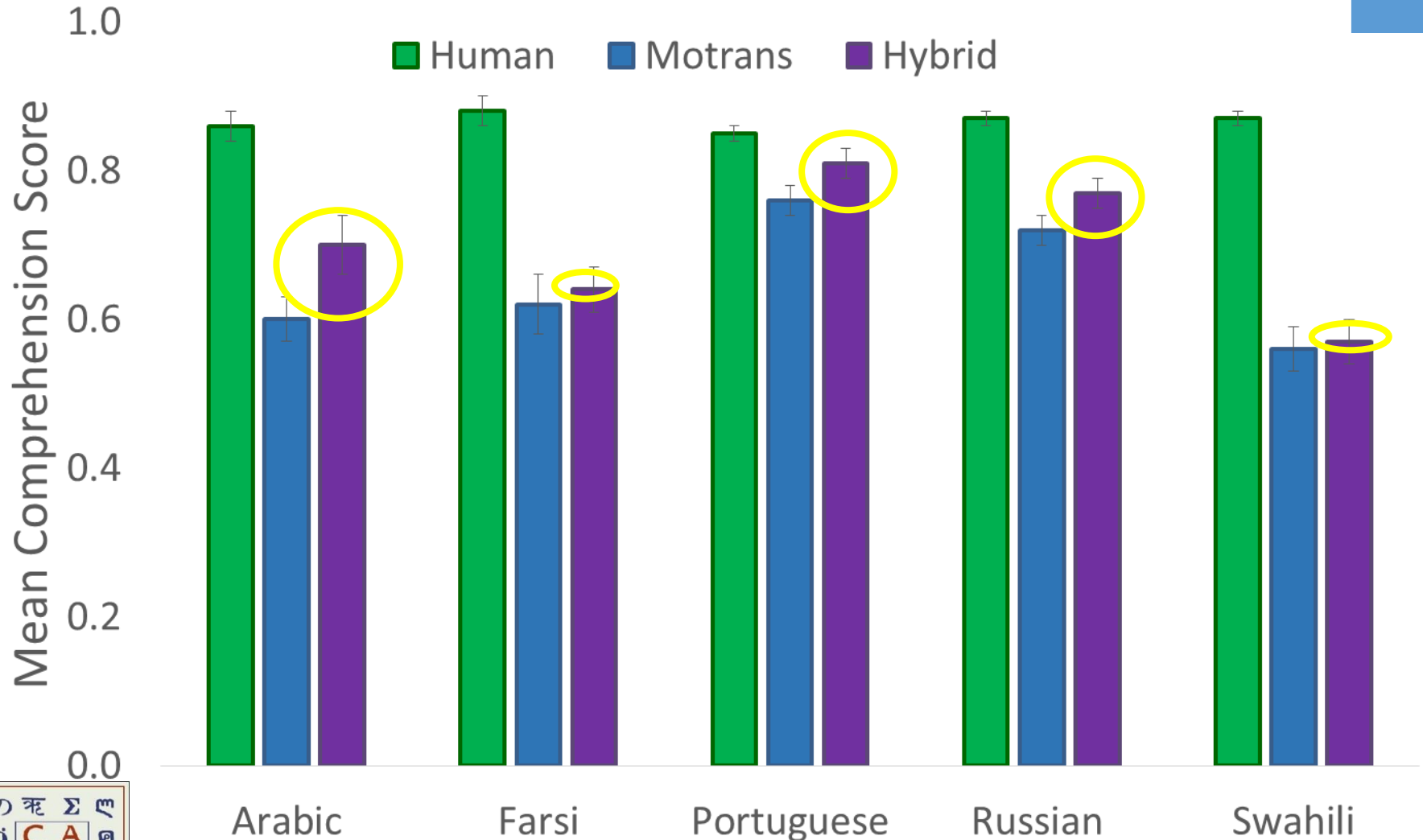
- Primary customers **don't** know language
  - Triage, filtering, selection
  - Free translators from spending time on low value material
- Secondary customers **do** know language
  - Gisting
  - Seed translation



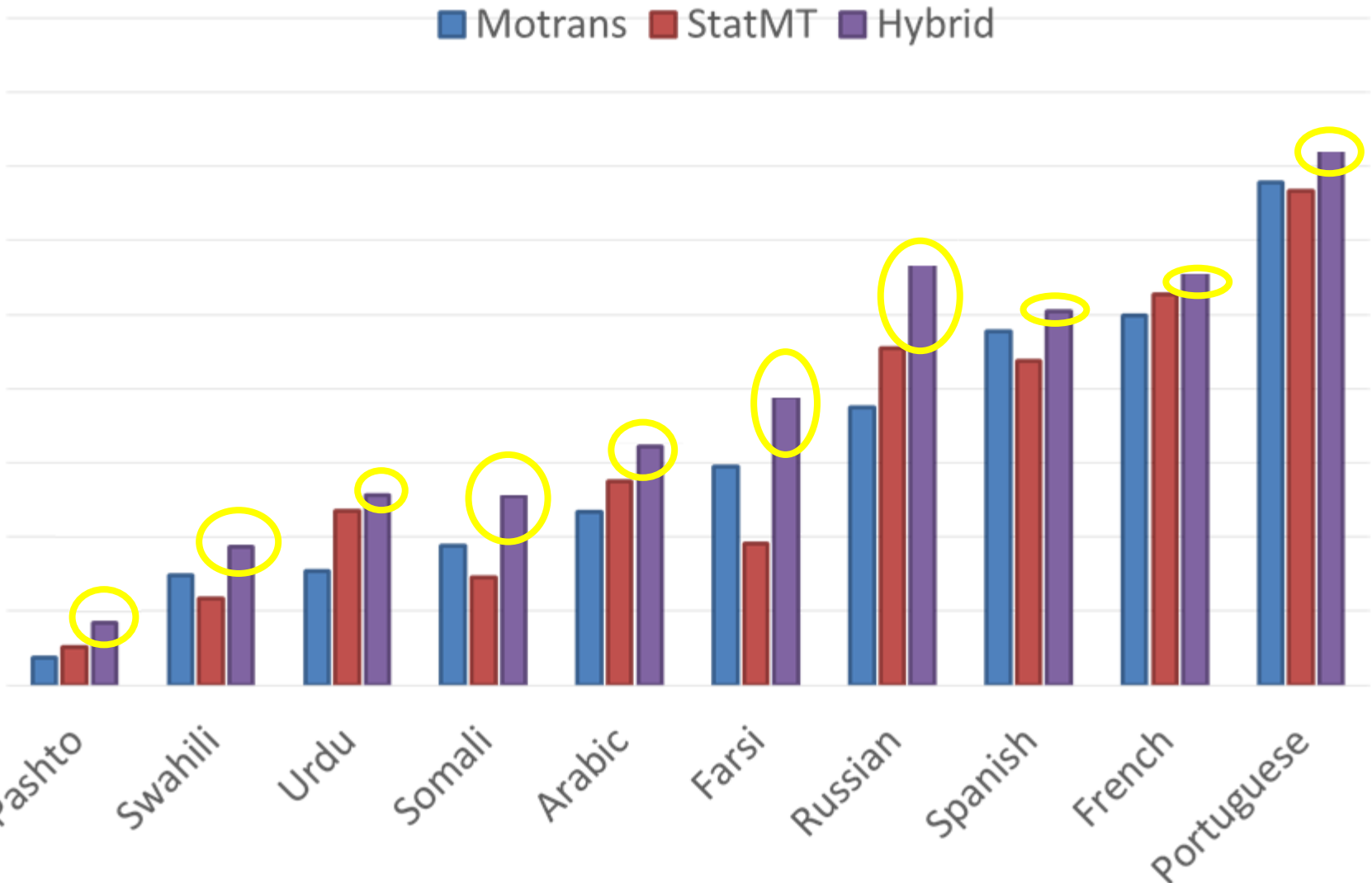
# Hybrid Approaches



# Human Comprehension Results

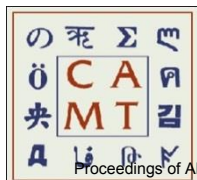


# BLEU Scores (In-domain)

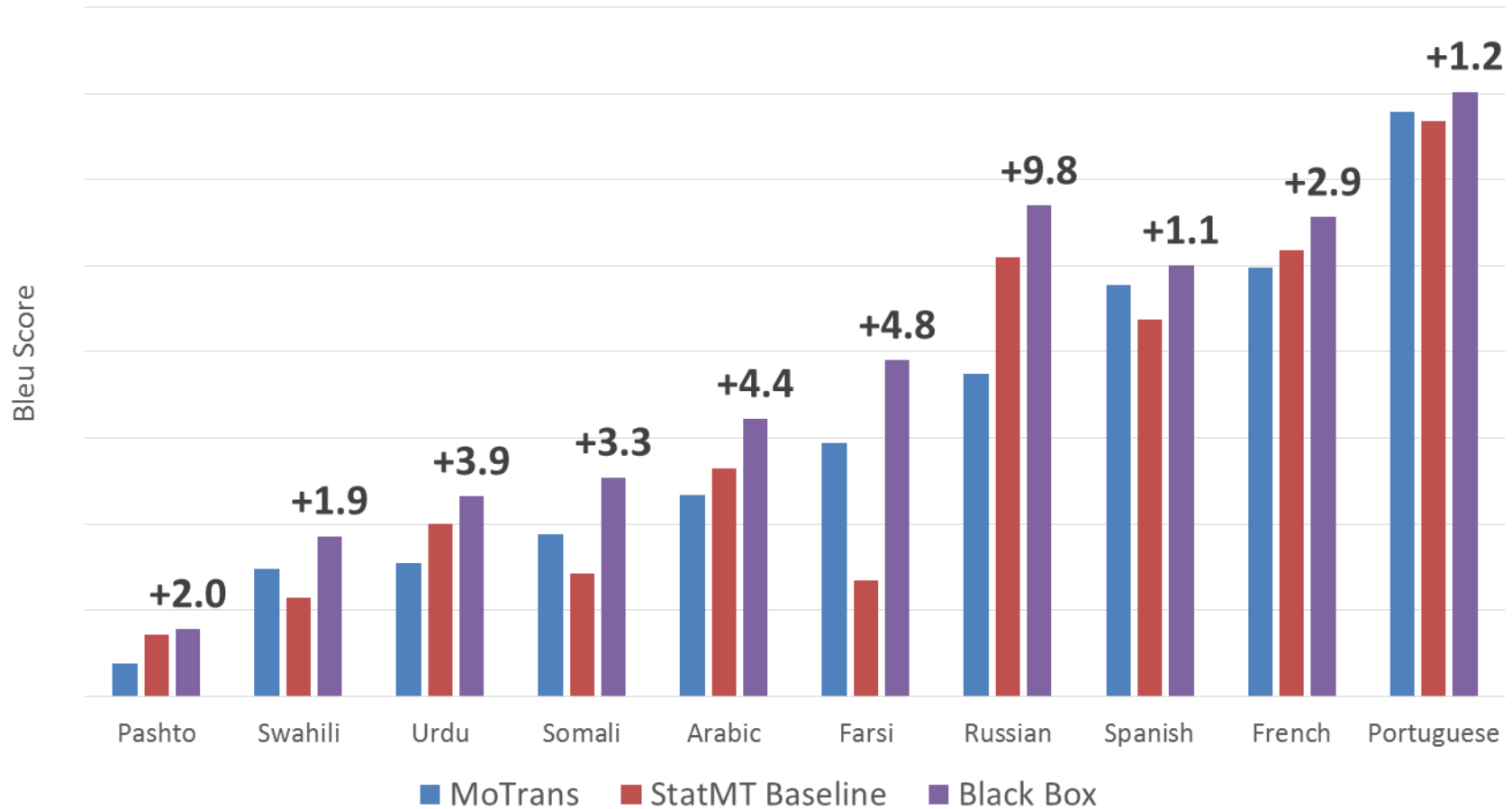


# Three Hybrid Approaches

<p><b>Black Box</b></p>	<p>Run Motrans on large text, build a regular StatMT model</p>
<p><b>Direct Conversion</b></p>	<p>Convert Motrans rules directly to StatMT phrase pairs</p>
<p><b>On Demand</b></p>	<p>StatMT system queries Motrans, incorporates its suggestions</p>



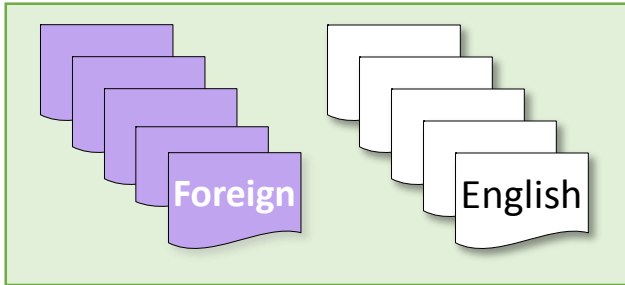
# Black Box results (in-domain)



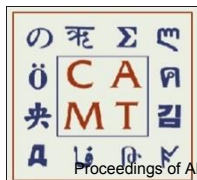
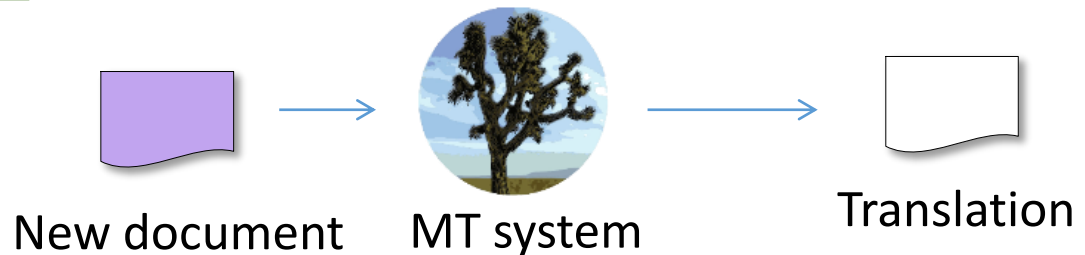
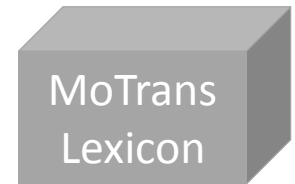
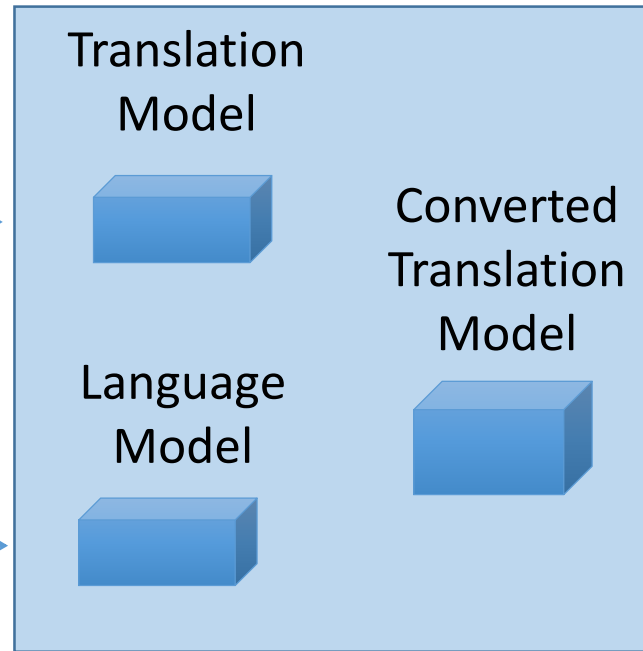
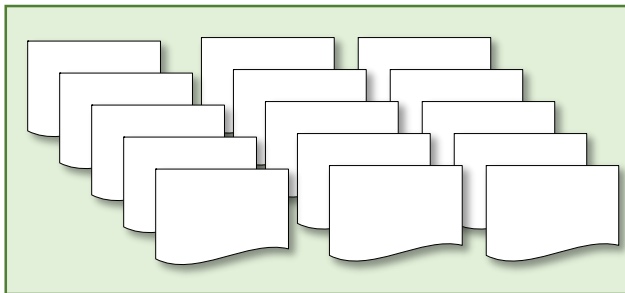


# Direct Conversion Approach

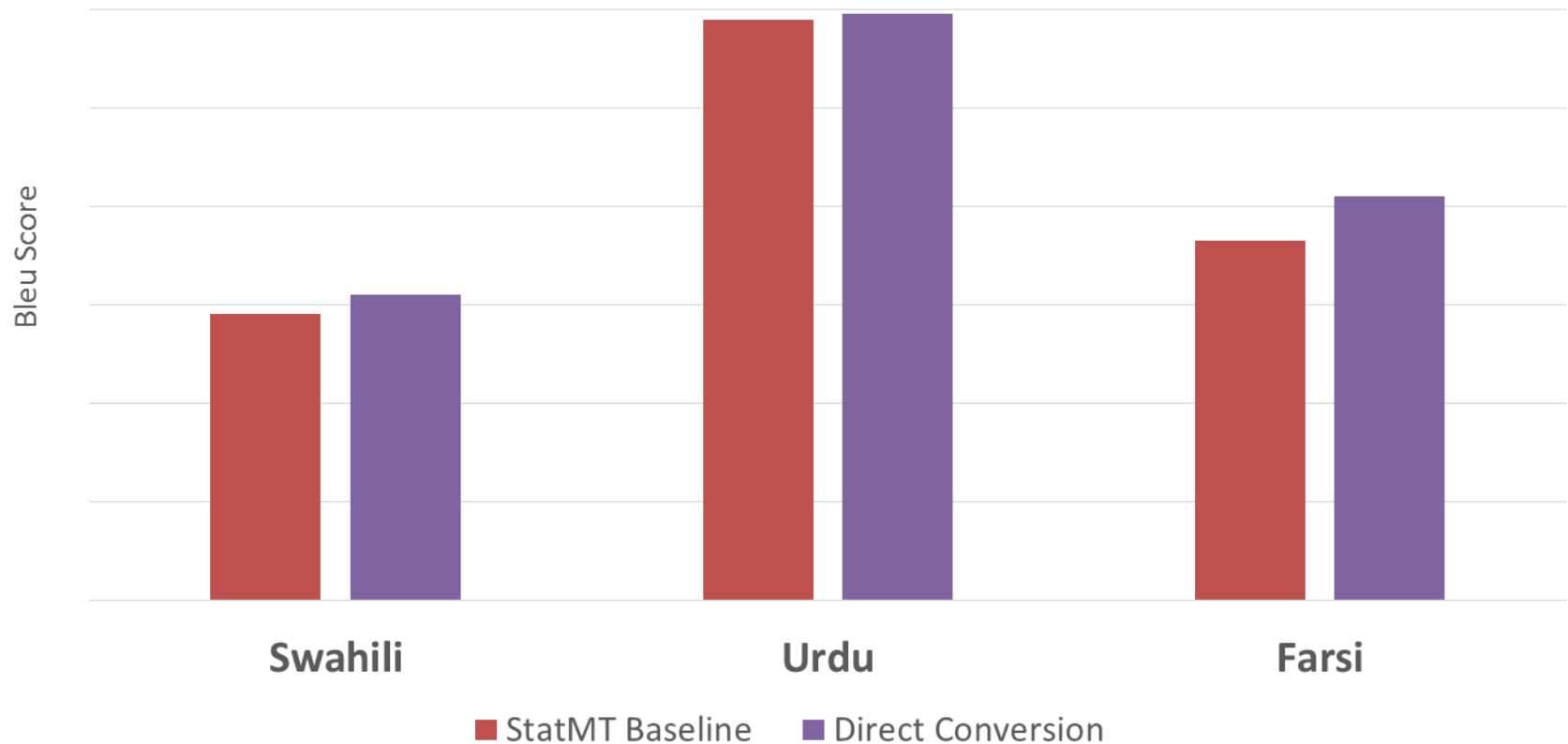
## Parallel sentences (bibtex)



## English text

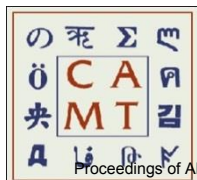


# Direct Conversion Results

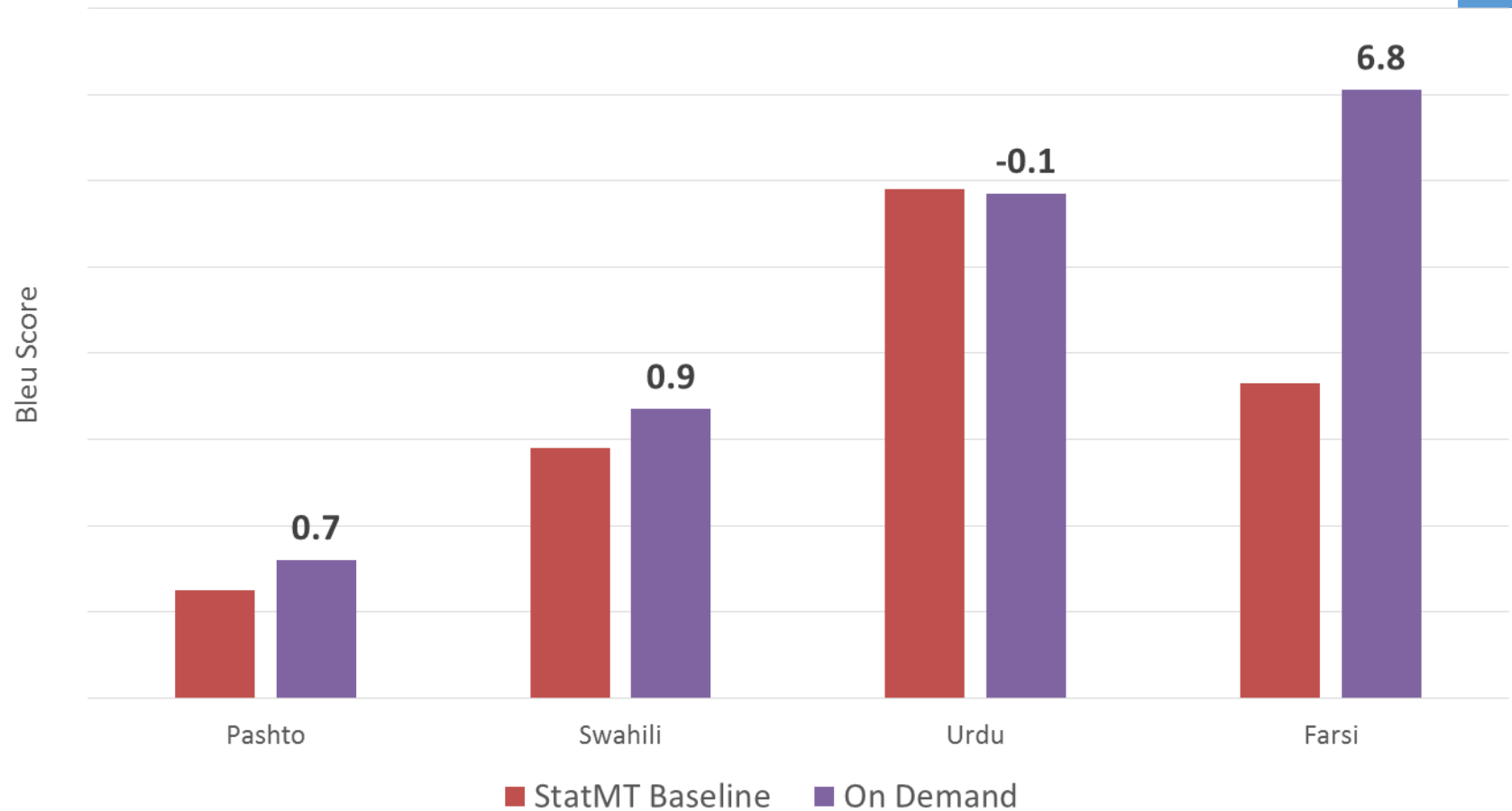


# Challenges of the Conversion Approach

- MoTrans rules have complex, unique syntax
  - Parsing requires in-depth knowledge of MoTrans
  - Some rule types not yet handled
- Not feasible to expand all rules
  - Some rules apply to full sentence, not phrases
  - Rule chaining creates exponential possibilities
- Cybertrans does pre- and post-processing that's not described in lexicon



# On Demand Results



# Combined Hybrid Results

