

# Solving Specific Content Challenges with Flexible Machine Translation

## MT Summit XV

Quinn Lam, Senior Program Manager, Machine Translation

November 2015





# What is Flexible Machine Translation?

# In the past ...

Foreign  
Language



## One-size-fits-all Machine Translation Platform

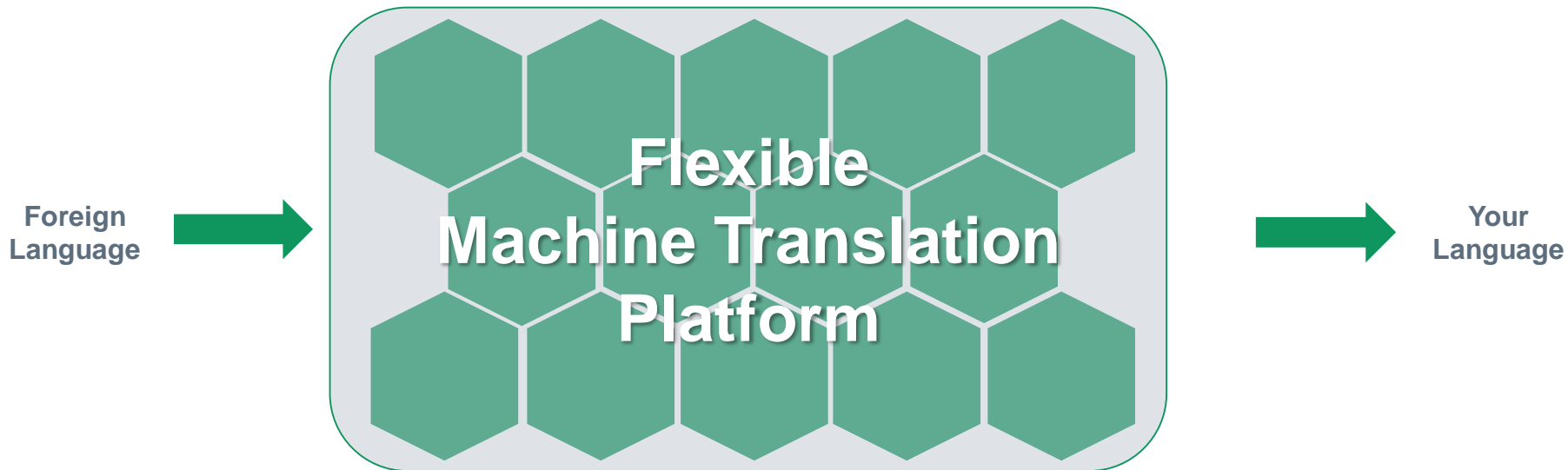


Your  
Language

Translation go through the decoding & training pipeline, regardless of

- The source & target language
- The content domain
- The translation use case

# Our approach



Modules could be added/removed/modified from the decoding & training pipeline, depending on

- The source & target language
- The content domain
- The translation use case

• etc

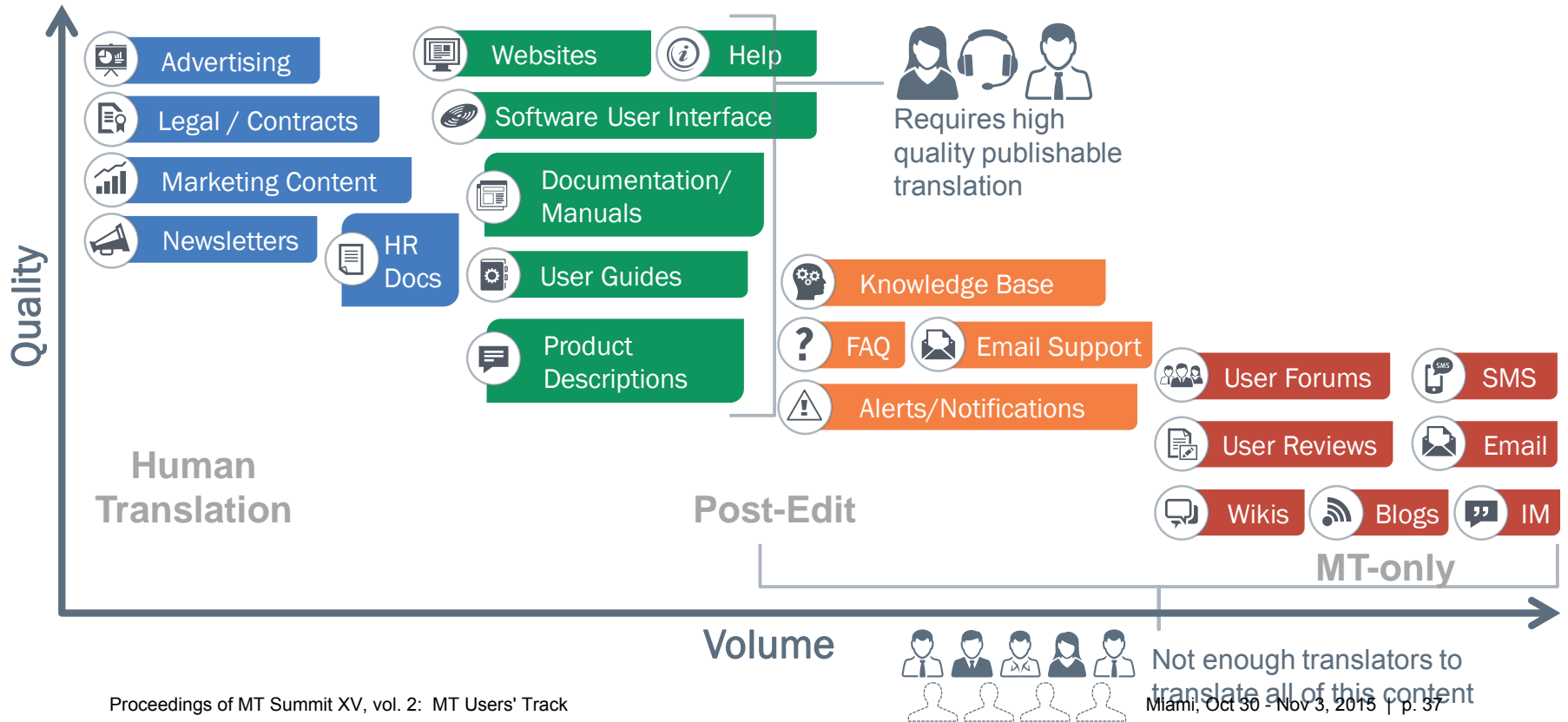
# Flexible Machine Translation in Action

- Specific challenges our customers encountered with MT
- Current MT shortfall in resolving those challenges
- SDL MT solution



# MT qualification

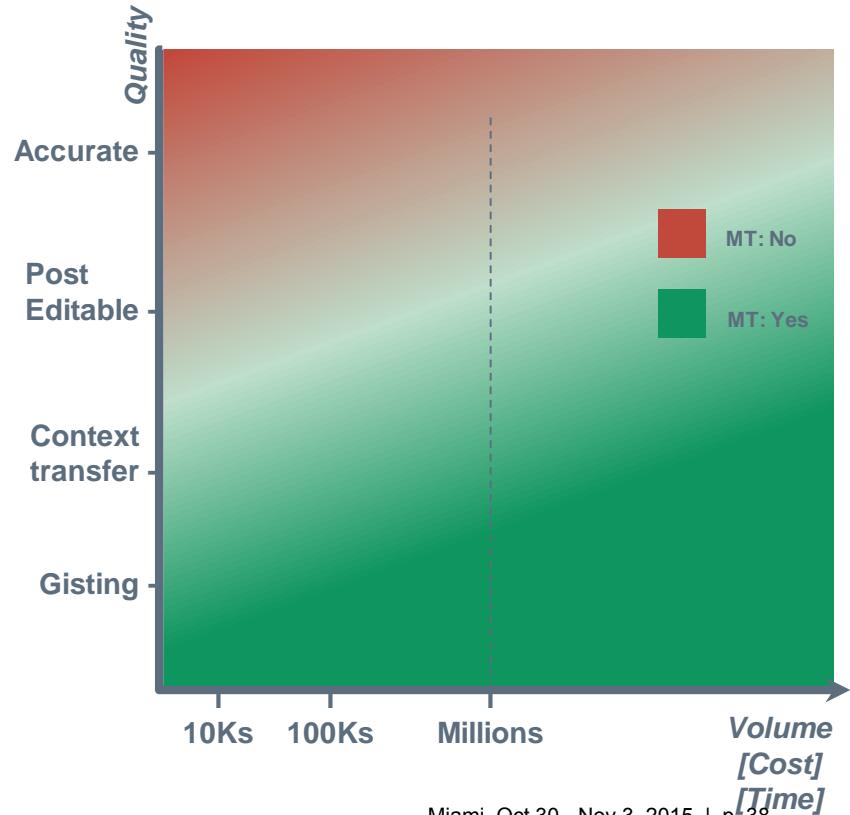
# MT Use-Cases



# Qualifying profiles for MT adoption


- While there are many important criteria, there are three key qualifiers for MT adoption:
  - 1. Speed:** Where content needs to be translated at a pace that humans can not match.
  - 2. Volume:** Where volume exceeds what can reasonably be accomplished (time and cost) by humans.
  - 3. Quality:** Ability to produce translations at a compelling quality. MT does not deliver perfect translations ("perfect" is subjective), but translation that are actionable.

Proceedings of MT Summit XV, vol. 2: MT Users' Track



Miami, Oct 30 - Nov 3, 2015 | p. 38





# Improving Number Translation Accuracy

# Customer A

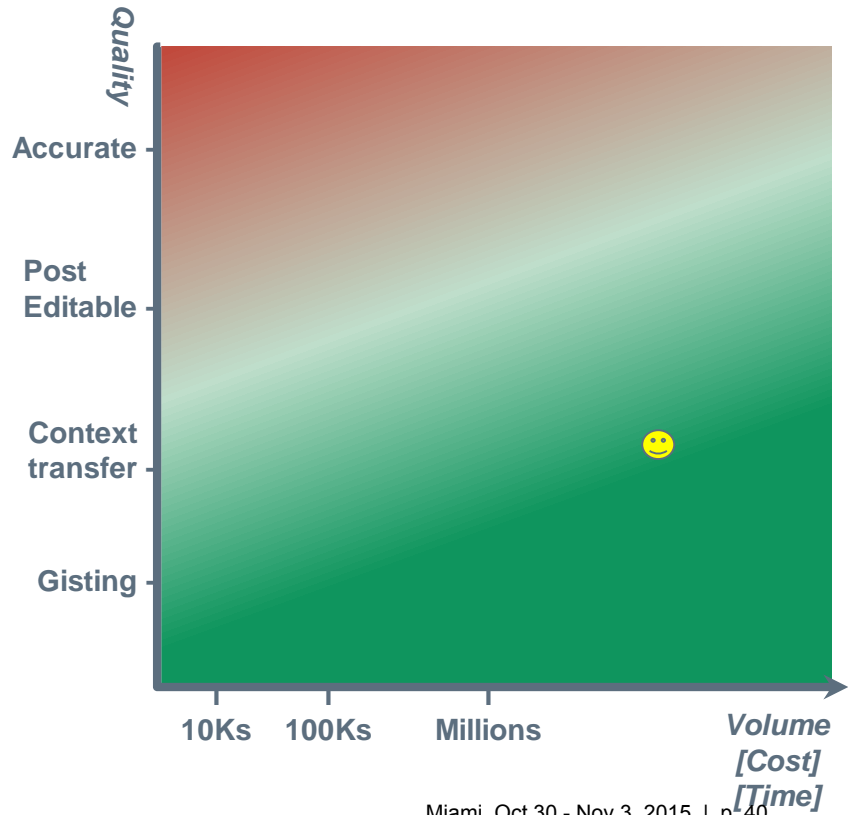
**Domain:** Financial

**Use-Case:** Analysis of financial news and press releases

**MT Solution:** Custom MT engines using customer's domain specific-data

- Higher Bleu scores when compared to baseline engines
- Higher Likert (Human Reviewer) scores through blind evaluation

Proceedings of MT Summit XV, vol. 2: MT Users' Track



Miami, Oct 30 - Nov 3, 2015 | p. 40

# Examples of Number Variations

	IQ 2008	IQ 2009		
<b>FAKTORZY</b>	obroty w mln PLN	obroty w mln PLN	% udział w rynku	% rok do roku
<b>ING Commercial Finance</b>	3127.0	2118.0	30.26%	-32.3%
<b>Pekao Faktoring</b>	1275.0	1104.0	15.78%	-13.4%
<b>Coface Poland Factoring Sp. z o.o.</b>	548.1	954.0	13.63%	74.0%
<b>Polfactor</b>	775.0	782.2	11.18%	0.9%

Aktualne notowania NBP (2015-08-27)				Przed miesiącem (2015-07-27)		
Waluta	Kurs ( PLN )	Zmiana		Waluta	Kurs ( PLN )	Zmiana mies.
1 EUR	4,2255	- 0.3 %	↓	1 EUR	4,1495	+ 1.83 % ↑
1 CHF	3,9260	+ 0.2 %	↑	1 CHF	3,9150	+ 0.28 % ↑

Nazwa	Otwarcie dnia	Kurs	Zmiana	Zmiana %	Minimum	Maksimum	Ostatnia zmiana
AUD / CAD	0,9477	0,9477	↑+0,0001	↑+0,01%	0,9429	0,9493	21:14
AUD / CHF	0,6802	0,6899	↑+0,0098	↑+1,44%	0,6760	0,6931	21:14
AUD / JPY	85,6900	86,3470	↑+0,6570	↑+0,77%	85,0750	86,9400	21:14
AUD / NZD	1,1046	1,1094	↑+0,0048	↑+0,43%	1,1015	1,1099	21:14

**Challenge for MT:** Numbers, currencies, dates, etc. need to be translated with high accuracy and consistency

- Dates need to stay in original language format
- Numbers need to be in original language format
- Currency symbols need to be correct
- Negative losses (-) and positive gains (+) need to be respected

Financial news are “digits heavy”

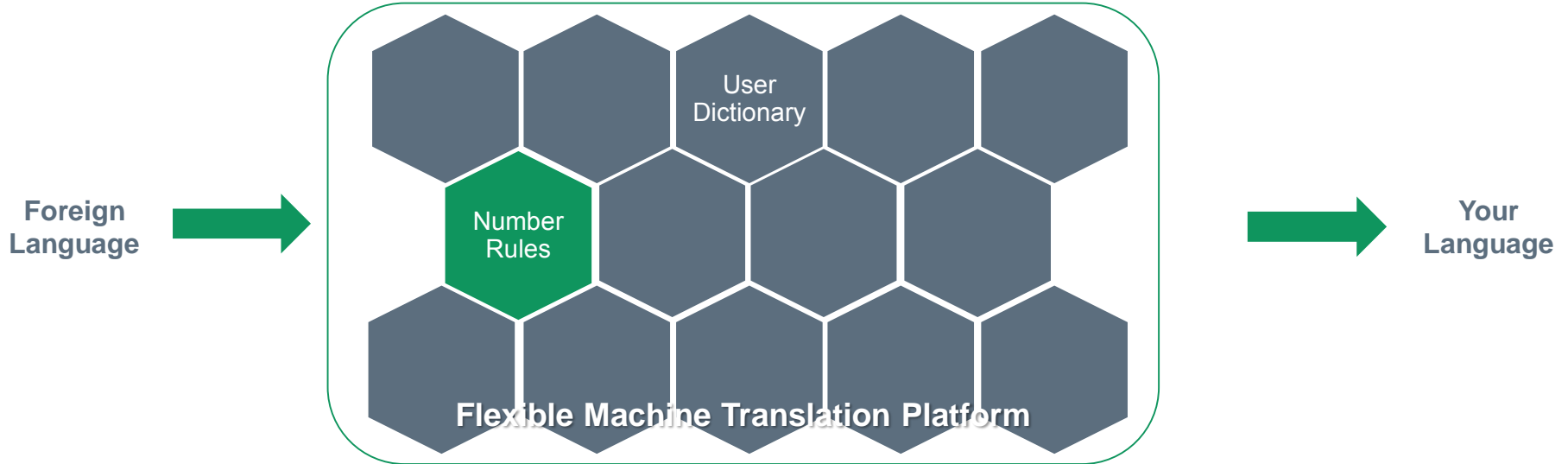
# Number Translation Issues

With use, customer found “surprising” issues related to how some numbers were translated

The statistical nature of SMT creates unpredictable number translation when SMT engine had learned incorrect phrase-pairs with numbers

Original Text	Output from custom Engine
(2)% (7)%	(2)·per·cent 7·%
827 (937) (71)	827 (937·WERE·closed) (71)
-110 -142 -157	-110 -142·With -157

# Solution with Flexible MT Platform



- 109/109 Number-related issues fixed
- Number Rules can be added as new pattern of issues occur
- No retraining of customer engines required
- Number Rules are compatible with user's dictionary usage



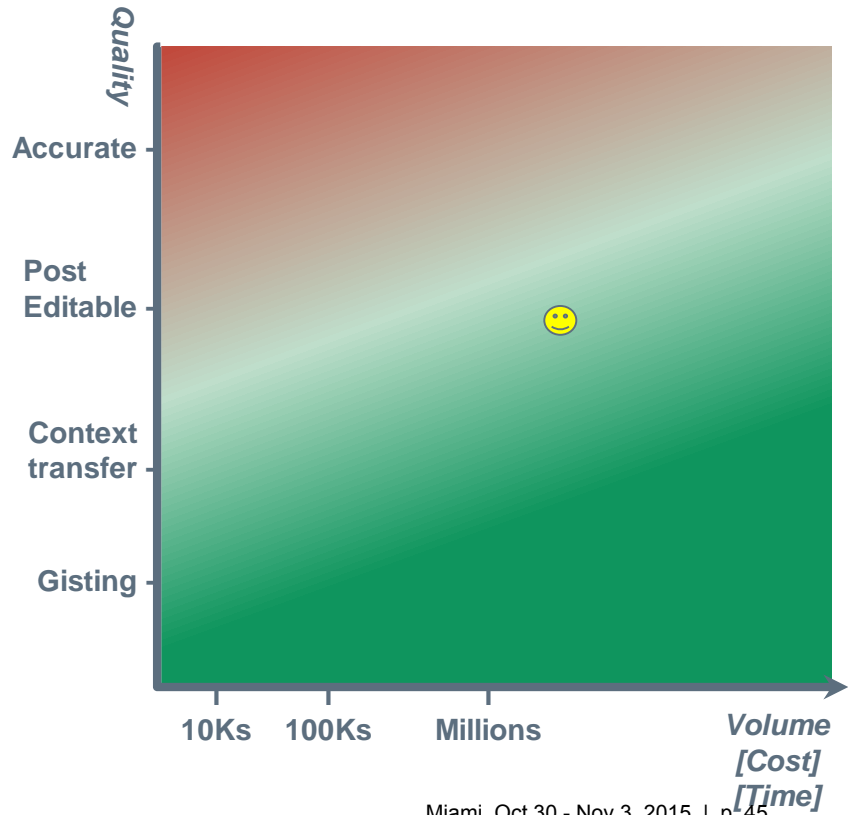
# Translating Unstructured Addresses to Structured Form

# Customer B

**Domain:** Mass Media

**Use-Case:** Detection and analysis of organized events

**MT Solution:** Generic MT engines to translate externally generated press releases and event details



# Examples of South Korea Addresses

South Korea Addresses	Romanized Addresses
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852
손양면 전사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun

**Challenge for MT:** Not enough “good” data for the MT engines to learn from.

- South Korea addressing systems continue to reform overtime
- Official address system is not the most common address system used



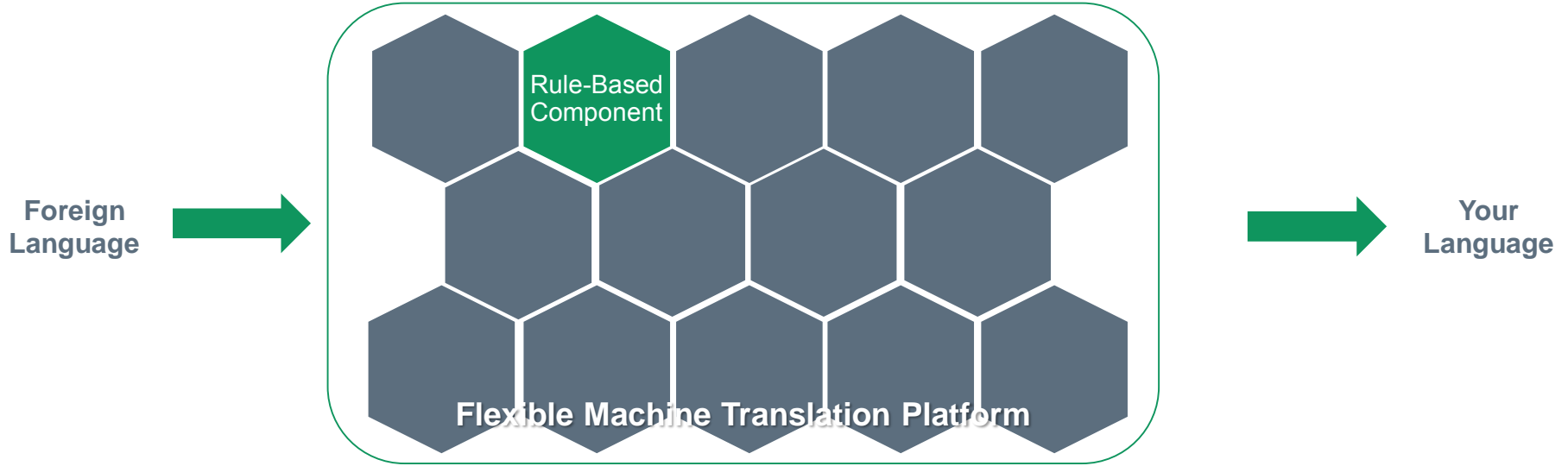
# Korea Addresses Through MT

Korea Addresses	Romanized Addresses	Baseline MT Engine
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014	Dong Geumcheon-gu, Seoul 153-014 378
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050	152-050 Seoul Guro-dong, 1128-1
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul	The Sangam, Mapo-gu in Seoul, 1587
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852	215-852 mulchiri in Gangwon Province on Mt. Yang Yang 16-3
손양면 선사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun	International Seonsayujeogro 678

# Solution with Flexible MT Platform

Korea Addresses	Romanized Addresses	Baseline MT Engine	Rule-based Component Enhancement
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014	Dong Geumcheon-gu, Seoul 153-014 378	153-014 378, siheung-daero, Geumcheon-gu in Seoul.
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050	152-050 Seoul Guro-dong, 1128-1	152-050 1128-1, guro-dong, Guro-gu in Seoul
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul	The Sangam, Mapo-gu in Seoul, 1587	1587, sangam-dong, Manangu in Seoul.
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852	215-852 mulchiri in Gangwon Province on Mt. Yang Yang 16-3	215-852 Gangwon-do16-3, mulchi-ri, Ganghyeon-myeon, Yangyang-gun
손양면 선사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun	International Seonsayujeogro 678	678, Seonsayujeok-ro, Sonyang-myeon

# Solution with Flexible MT Platform



Before:

- Exact match - 1465/26110 (5.6%)

After:

- Exact match – 8248/26110 (31.6%)

**463%**

improvement in  
exact translation



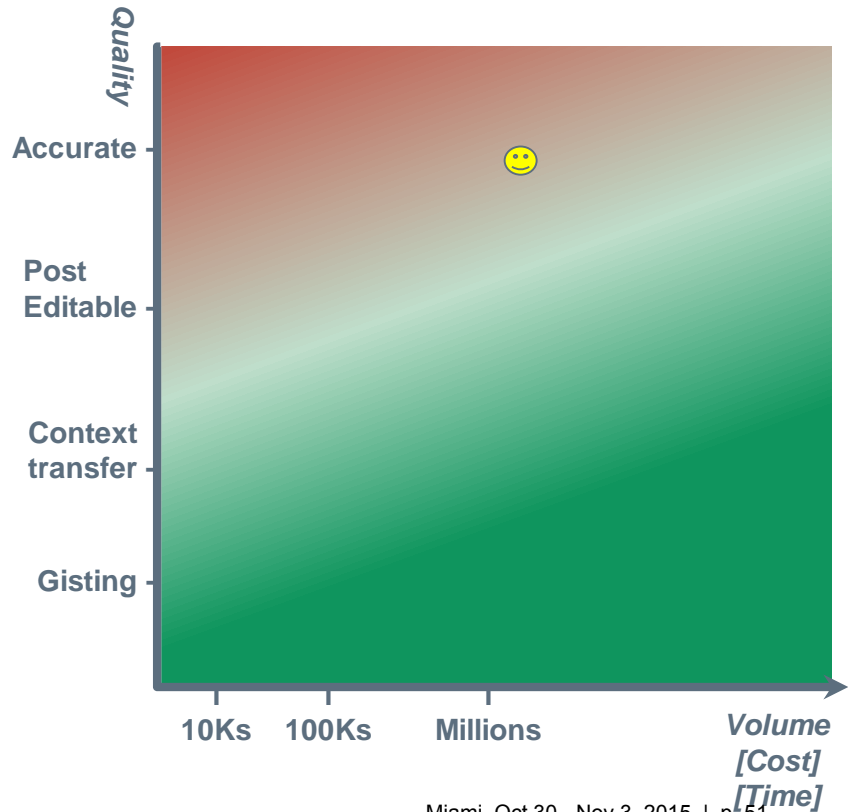
# Raising MT Quality of Dissimilar Syntax Language Pairs

# Customer C

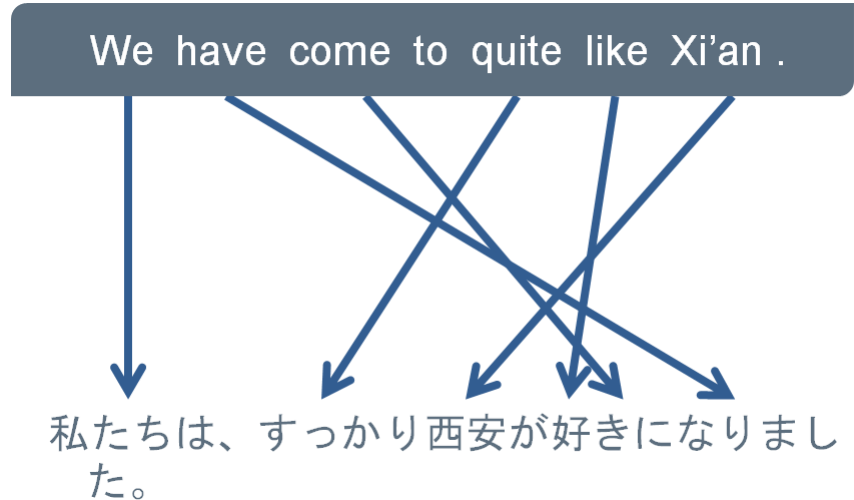
**Domain:** Language Service Provider (LSP)

**Use-Case:** Translation productivity

**MT Solution:** Generic Japanese <> English engines to be used in projects whenever possible



# Example of English to Japanese Translation



**Challenge for MT:** English and Japanese are syntactically very different

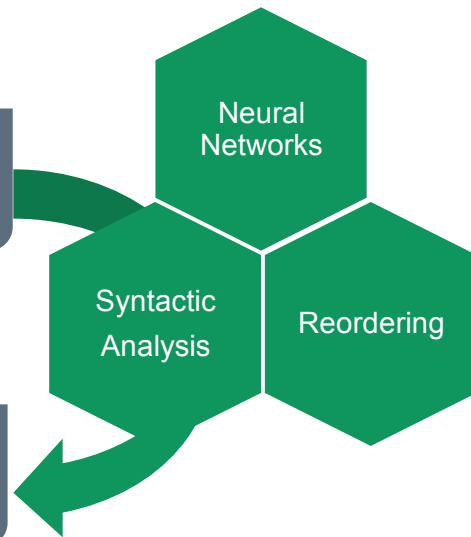
- Makes reordering of words and phrases to well formed sentences difficult for MT

# Solution with Flexible MT Platform

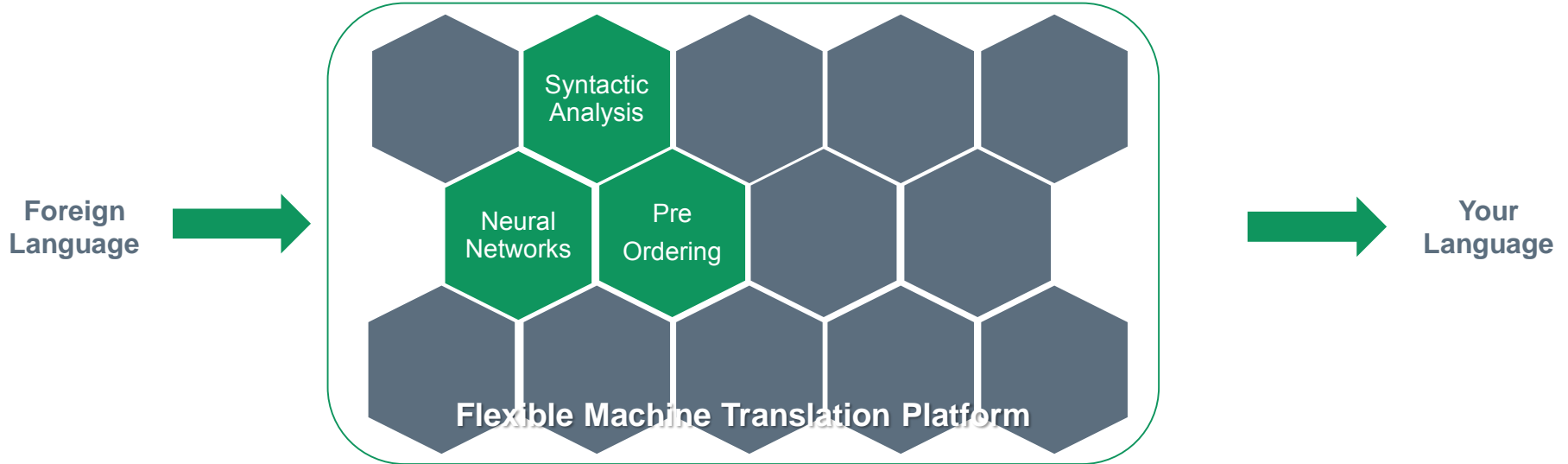
We have come to quite like Xi'an .

We quite Xi'an like to come have .

私たちは、すっかり西安が好きになりました。



# Solution with Flexible MT Platform



**Modules added to Eng<>Jpn engines during the training of the engines and are used during decoding time**

- **5% Bleu score improvement**
- **+0.08 Likert score improvement**





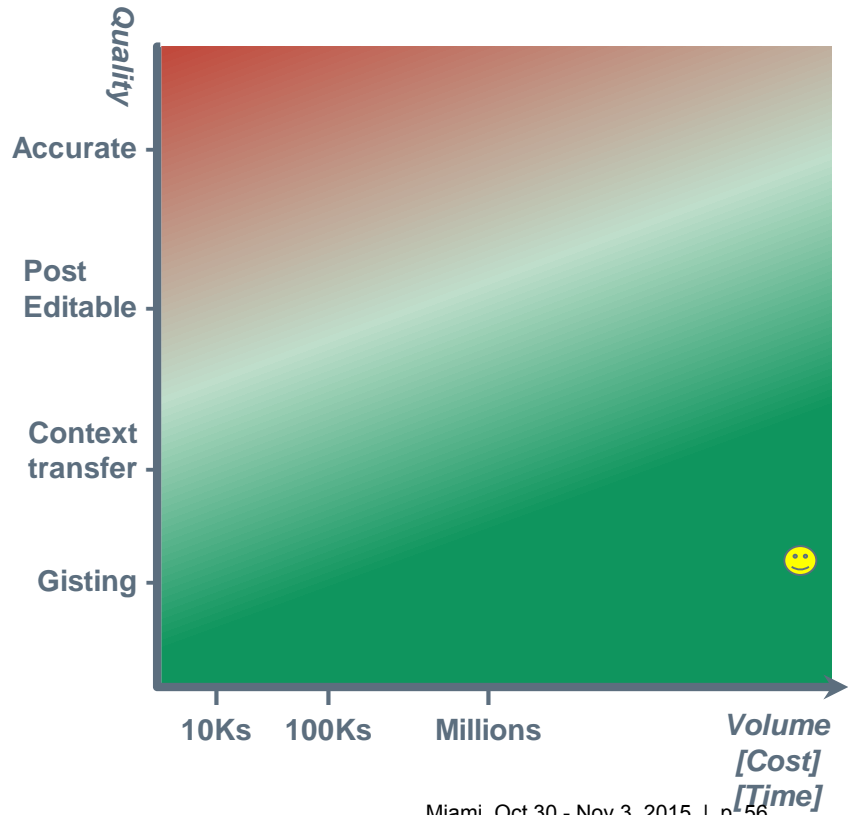
# Translating Social Media Morphed Content

# Customer D

**Domain:** Social Media

**Use-Case:** Sentiment analysis

**MT Solution:** High volume of user-generated informal text through customized MT engine (using domain specific data)



# Example of Informal Text

**Kareem** @kareem 8h  
ول كانت هده الحسن عطله افر مرح كثير المرفهين كانو حلوين وديمن وجدين روم سرفيس رتبو السرير بروعه والغرف كنت جد نزيفه #Mazagan  
Expand

**Jamal** @jamal 8h  
@kareem bess7a wel3afya 7abibi!  
Expand

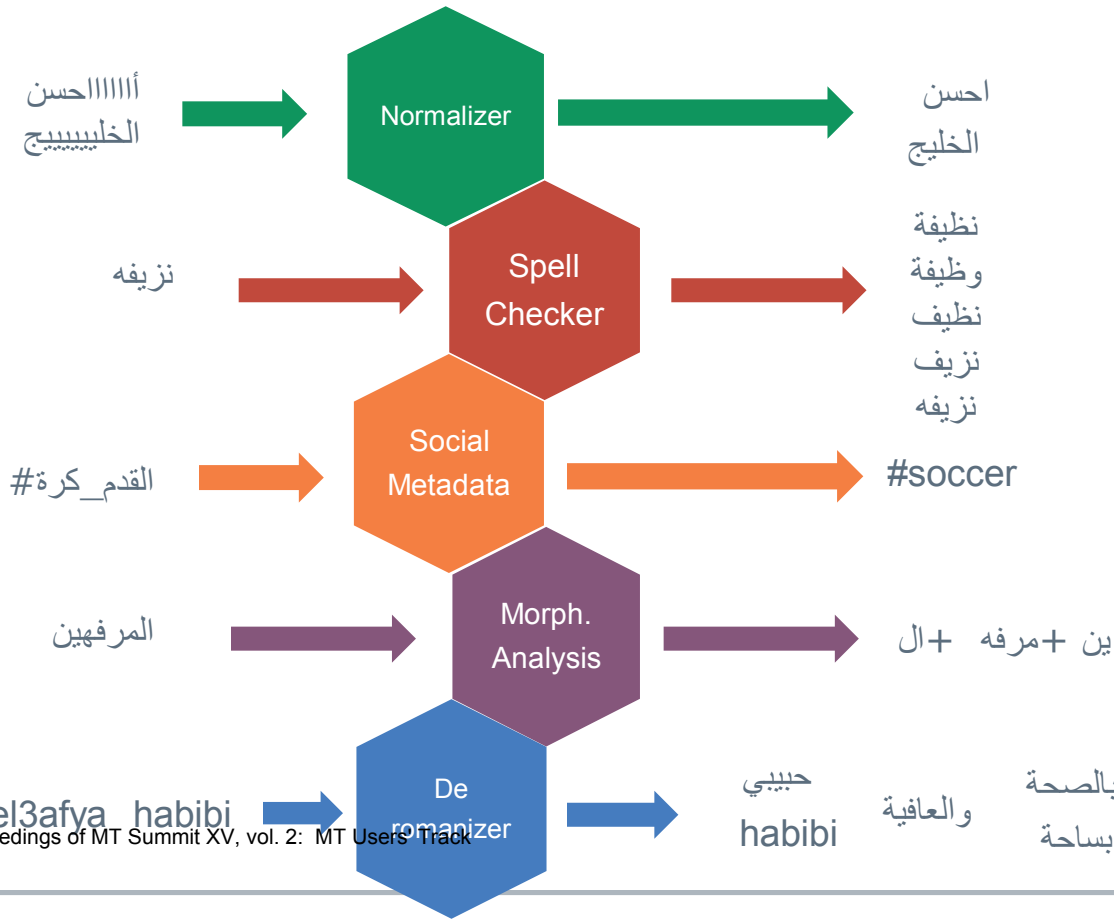
**Challenge for MT:** user-generated text found on social media are informal and short-handed writings.

Contains issues for MT such as

- **Character Repetition**

- **Spelling Errors**
- **Morphology**
- **Metadata**
- **Romanization**

# Solution with Flexible MT Platform



# Example of Romanized Arabic Translation

## Source

- la2a hia katir fi lakhbar.
- ma 3ajbanish kida. Lazim t3'iyer l3ounouane
- Enty habla ?
- Kalemni lama t3raf ezay tebatal teshtemni
- 3andy soda3 fi rassi... 5oshy namy badal chat. a7san lik Ah sa7

## Existing MT

- La2a hia katir Fi lakhbar.
- Ma 3ajbanish kida. lazim T3 (iyyer L3ounouane
- enty habla?
- kalemni Lama T3RAF ezay tebatal teshtemni
- 3Andy soda3 Fi rassi ... 5oshy namy badal Chat. A7San lik Ah SA7

# Solution with Flexible MT Platform

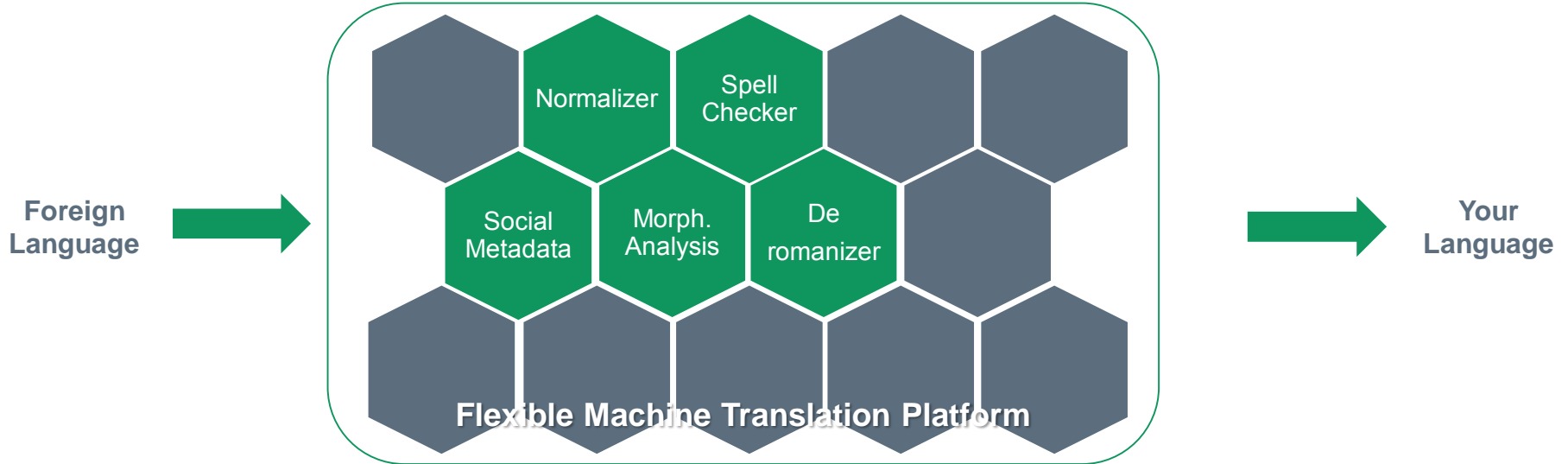
## Source

- la2a hia katir fi lakhbar.
- ma 3ajbanish kida. Lazim t3'iyer l3ounouane
- Enty habla ?
- Kalemni lama t3raf ezay tebatal teshtemni
- 3andy soda3 fi rassi... 5oshy namy badal chat. a7san lik Ah sa7

## Informal MT

- No, it is very much in the news.
- I don't like this. We must change the title
- Are you an idiot?
- Talk to me when you know how to stop insulting me
- I have a headache in my head. Go to sleep, instead of chat. It is better for you, Yes, sa7

# Solution with Flexible MT Platform



**Normalizer, Spell Checker, Social Metadata and Morphology Analysis modules:**

- 72% improvement in translated text

**Deromanizer module:**

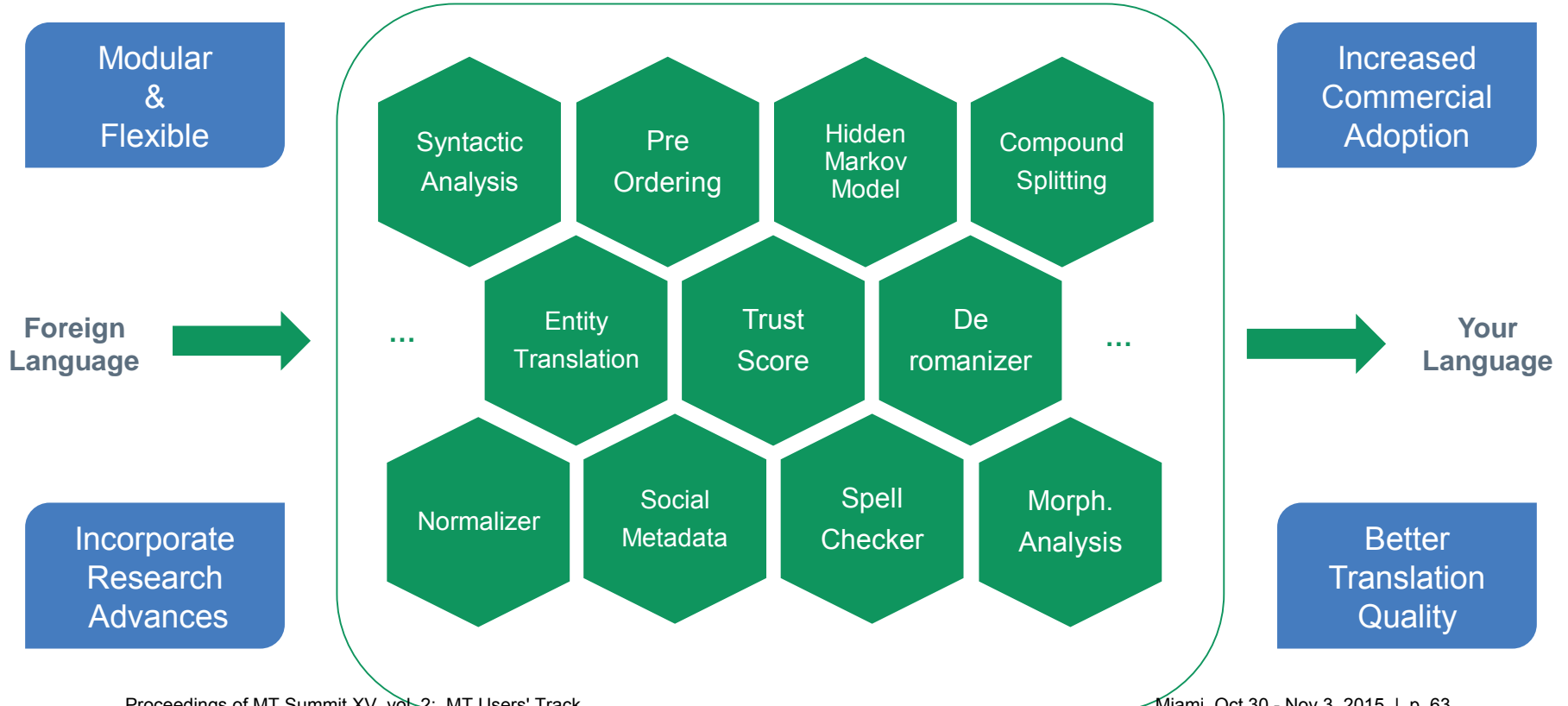
- No translated text to mostly translated text



# Flexible MT at a Glance



# Flexible Translation Architecture





## Global Customer Experience Management

Copyright © 2008-2014 SDL plc. All rights reserved. All company names, brand names, trademarks, service marks, images and logos are the property of their respective owners.

This presentation and its content are SDL confidential unless otherwise specified, and may not be copied, used or distributed except as authorised by SDL.