The effect of shallow segmentation on English-Tigrinya statistical machine translation

Yemane Tedla and Kazuhide Yamamoto
Nagaoka University of Technology
Japan

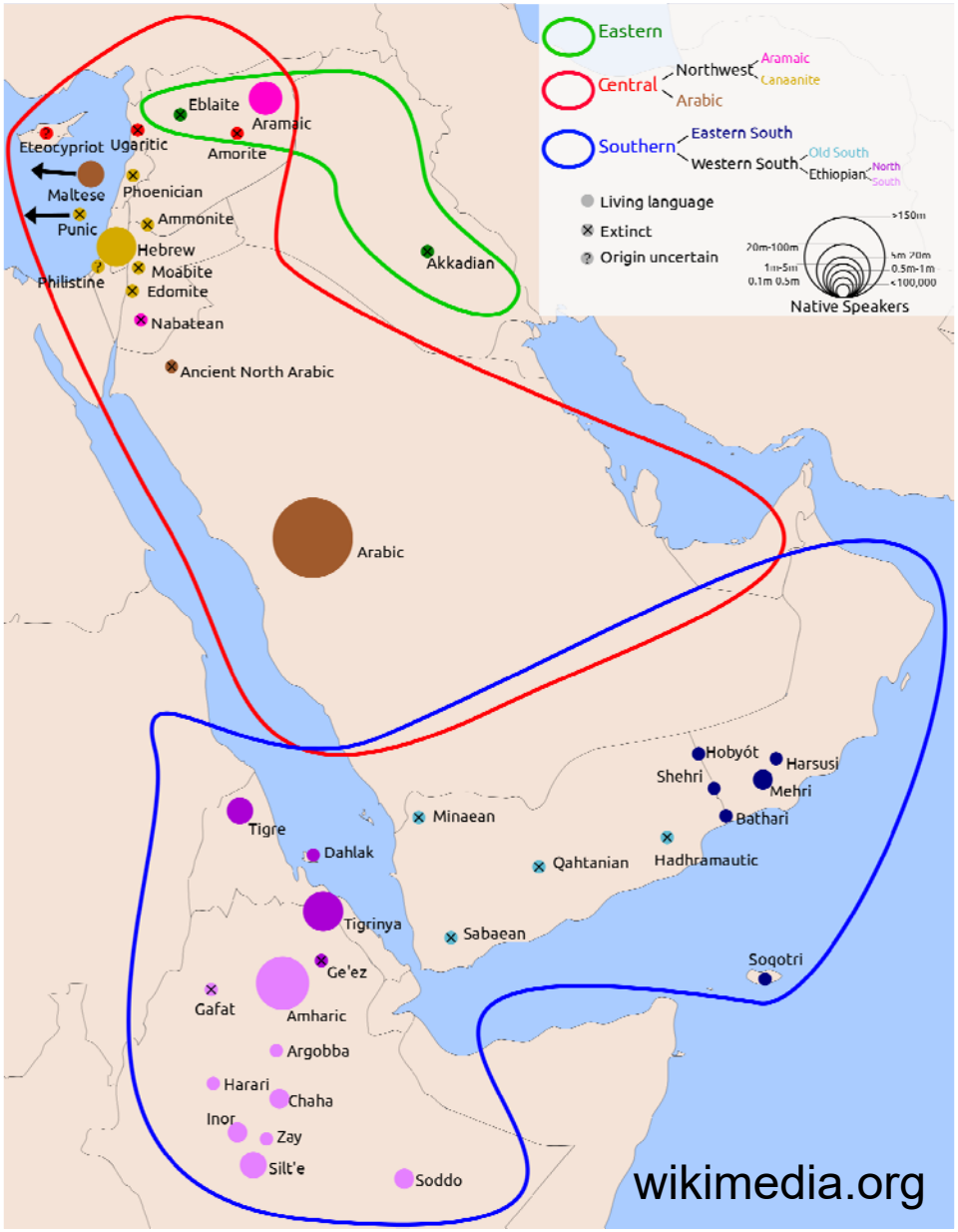
Tigrinya is native to Eritrea and Ethiopia

© Google maps

Population :
7 million +

Language Group

- Afro-asiatic
  - Semitic
    - Arabic
    - Hebrew
    - Tigrinya

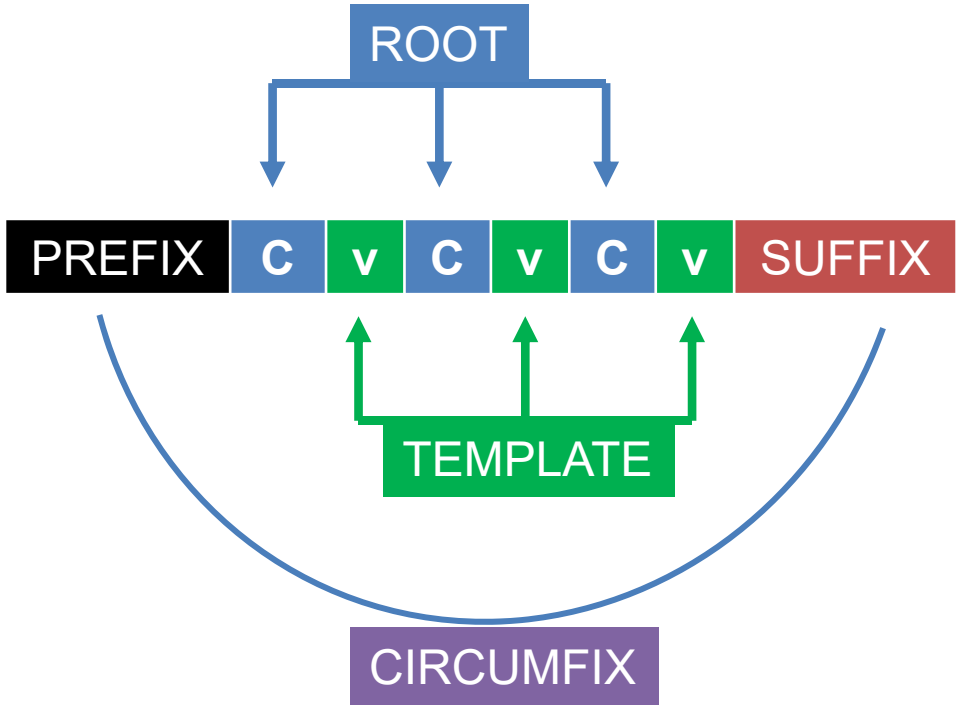


Tigrinya language

- Morphology
  - Root-template
- Highly inflected for
  - Gender, number, person
  - Aspect, mood, tense etc.
- Writing system – Ge'ez script
- Word order Subject-Object-Verb (S-O-V)
- Subject-verb agreement

ትግርኛ ምስ በዓል ቋንቋ ዓረብ ፡ ዕብራይስጥን ኣምሓርኛን ዝምደብ ሴማዊ ቋንቋ እዩ። እዞም ቋንቋታት ኣብ ትሕቲ ስድራ-ቤት ኣፍሮ-ኤስያውያን ቋንቋታት ይጥርነፉ። ቋንቋ ማልታ፡ ኣራማይስጢ፡ ሱርስት (ሶርያ)፡ ትግረ ከምኡውን ሕጂ ኣብ ኣገልግሎት ቤተክርስትያን ጥራይ ተሓጺሩ ዝርከብ ጥንታዊ ግእዝ ውን ካልኦት ሴማውያን ቋንቋታት እዮም። ትግርኛ ኣብ ኤርትራን ሰሜን ኢትዮጵያን (ትግራይ) ይዝውተር። ናይ ክልቲኡ ሃገራት ብዝሒ ተዛረብቲ እዚ ቋንቋ እዚ ልዕሊ 7 ሚልዮን ከም ዝበጽሕ ይግመት። ኣብ ብዙሓት ሃገራት ኤውሮጳ፡ ኣሜሪካ፡ኣውስትራልያ፡ ኣፍሪቃ ከምኡውን እስራኤል ብዙሓት ተዛረብቲ ትግርኛ ኣብ ስደት ይነብሩ።

Tigrinya morphology (root-template morphology)



C – CONSONANT
V – VOWEL

Tigrinya morphology (root-template morphology example)

Word	sebere	[He] broke
template	-e-e-e-	TAM inflection occurs by vowel alterations
root	s-b-r	Consonants represent the concept ‘to break’
prefixing	**inte**sebere	If [he] broke
suffixing	seberna	[we] broke
Infixing	sebira	[she] broke
circumfixing	ayseberen	[he] did not break

[pronoun] – the pronoun is not explicitly stated but inferred from the verb since there is verb – subject agreement in Tigrinya

Tigrinya Natural Language Processing

- **NO** publicly available corpus **until 2015**
- **Now available text corpus are:**
  - POS tagged corpus (our research)
    - *eng.jnlp.org/yemane/ntigcorpus*
  - text condordancing (Habit project)
    - ***habit-project.eu/wiki/InterimResults***
- **Few** other works on Tigrinya NLP
  - Morphological analyzer and generator (2011)
  - Stemmer for Tigrinya (2011,2013)
  - Tigrinya Search engine (2013) at Nagaoka University of Technology
  - Some Input Method Editors (IME)
  - English-Tigrinya electronic dictionaries

Our Tigrinya NLP research

- Text corpus construction
- Manual POS tagging of around 72K words
  - available at ***eng.jnlp.org/yemane/ntigcorpus***
- Part-of-speech tagger with Tigrinya morphological patterns
- Morphological segmentation and statistical machine translation

Tigrinya morphological segmentation

Tigrinya - InItezeyIHatetIkayo

English - if you did not ask him

Word alignment is difficult without segmentation

Morphological segmentation

Token (ti) InItezeylHatetIkayo

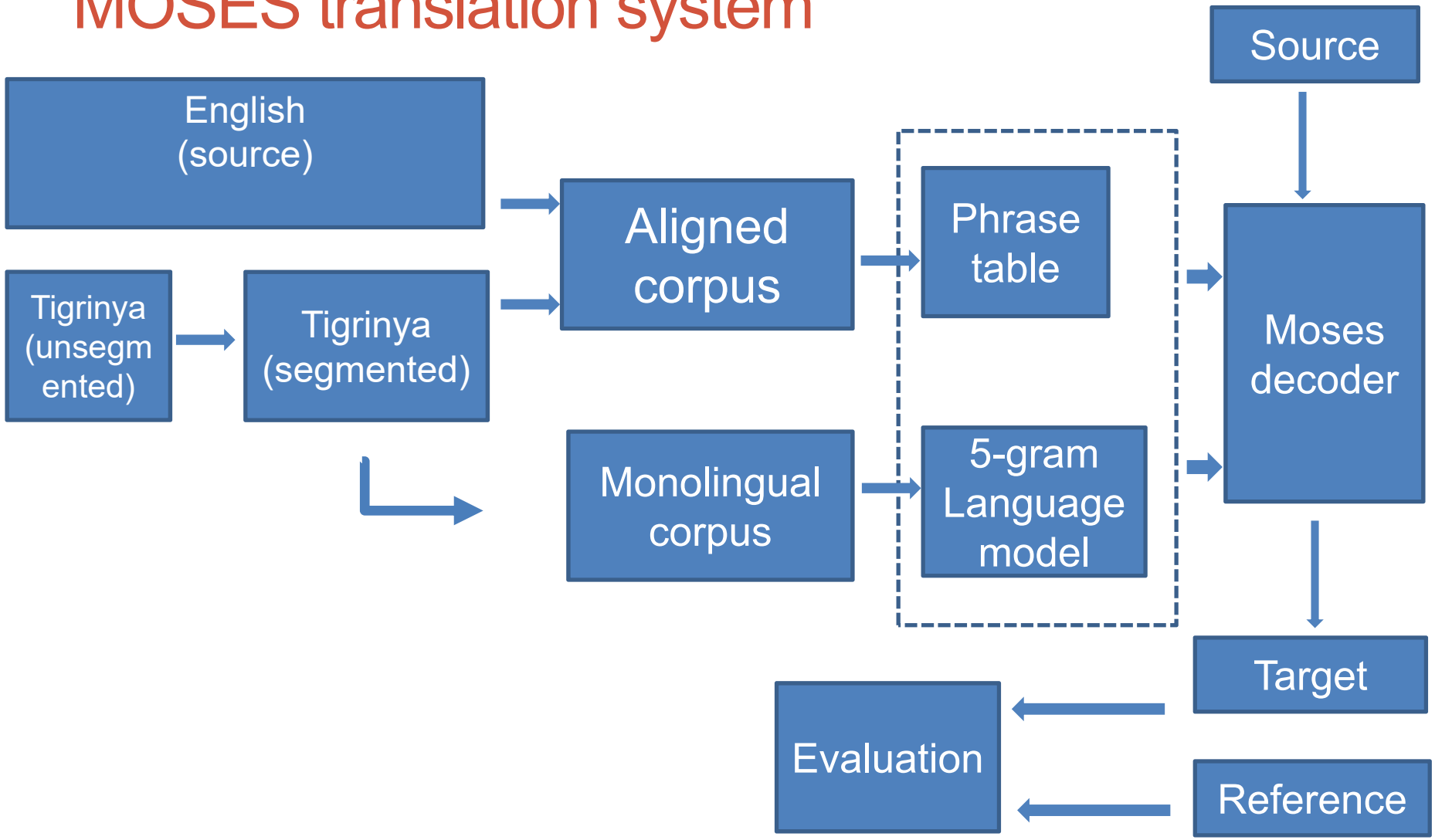
Gloss(en) If you did not ask him

fine segm. InIte zeyl HatetI ka yo

Shallow segm. **(our research)** InItezeyl HatetI kayo

Prefix Stem suffix

Experiment - MOSES translation system



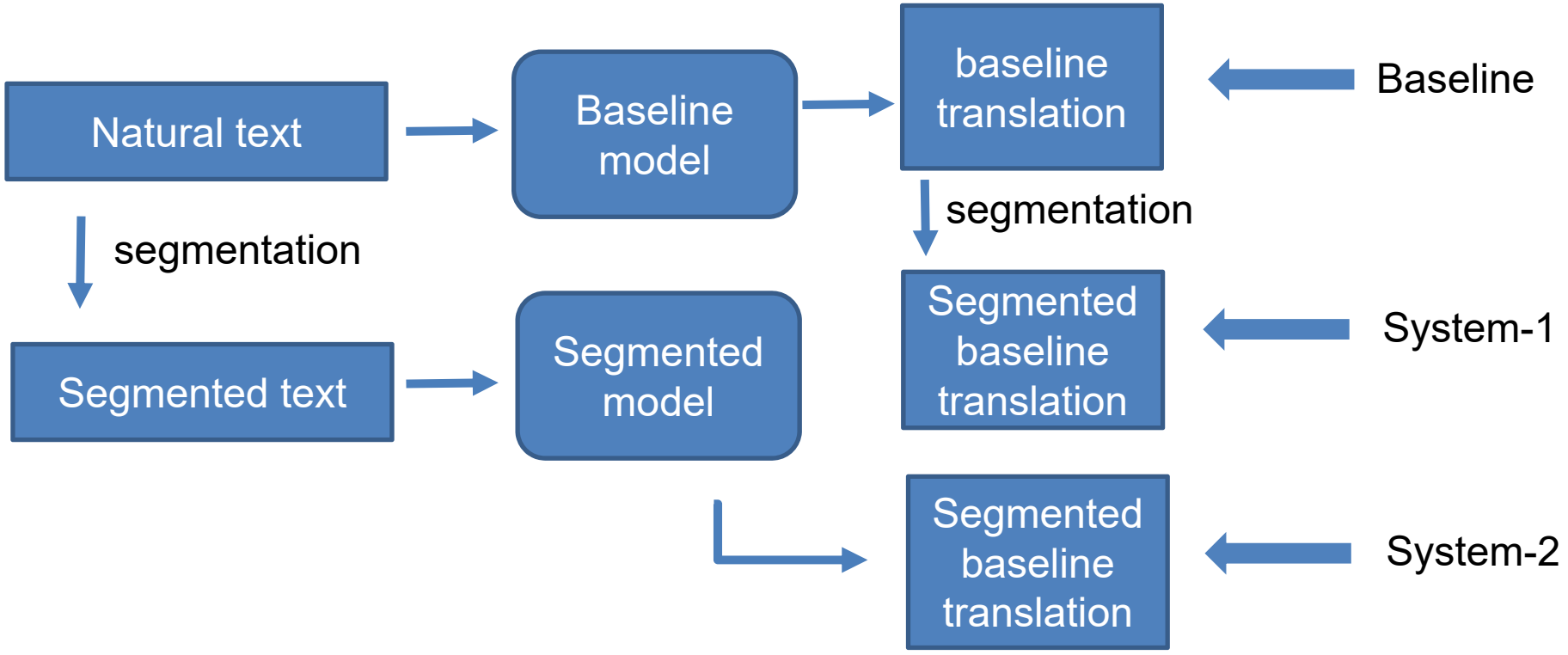
Parallel Corpus

- The only parallel corpus available is the Bible (English-Tigrinya translations) (***geezexperience.com***)
- Preprocessing the Bible
  - English verses are found sequentially
    - (1,2,3,4,5,….)
  - In Tigrinya verses are frequently found combined
    - (1,2,**3-4**,5…)
  - We joined the corresponding English verses for proper alignment

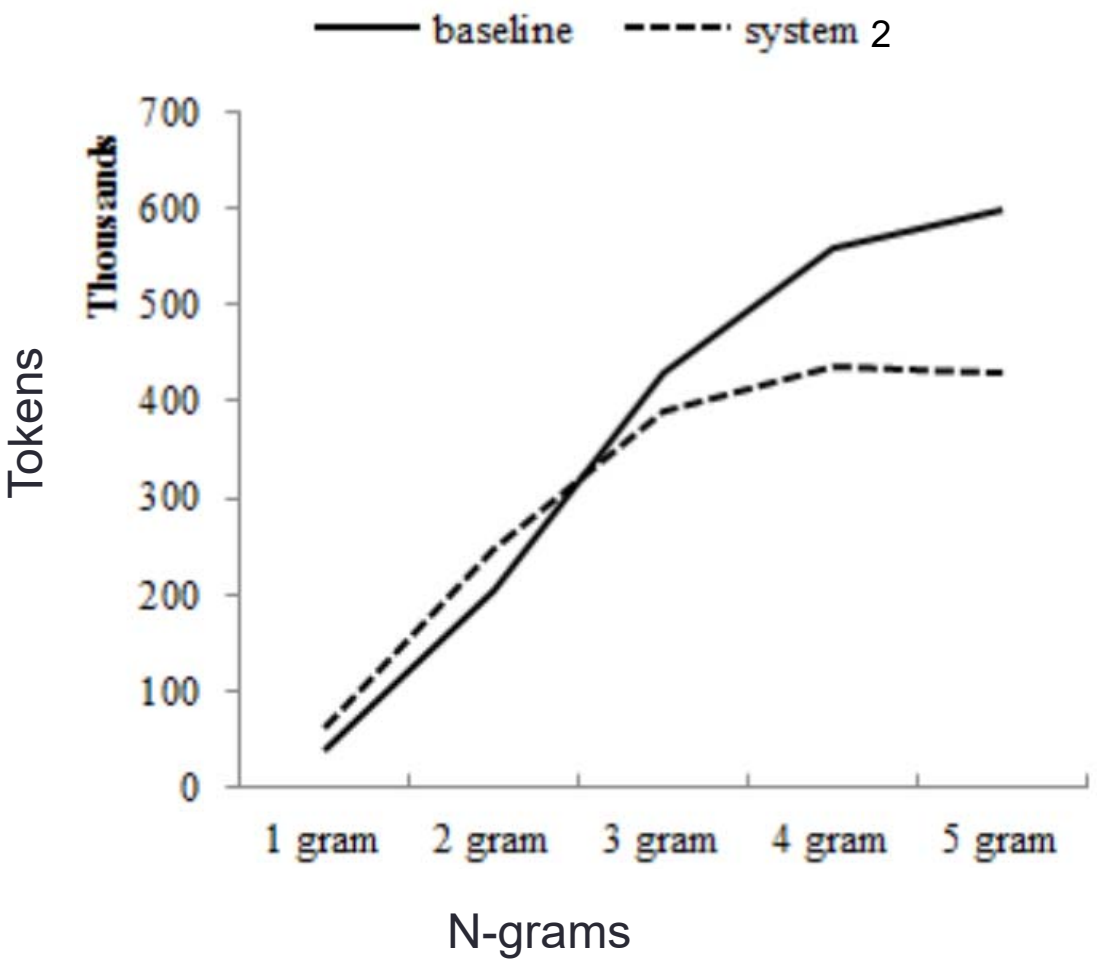
Data	Verses	English tokens	Tigrinya tokens unsegmented	Tigrinya tokens segmented
Training	31,279	938,837	584,318	837,675
Test	1,000	31,994	20,042	28,808
Tuning	970	31,383	19,624	28,254

Models and evaluation

- Baseline system (natural text model)
- System-1 (baseline translations segmented to suit comparison with System-2)
- System 2 (segmented text model)



Effect in language model



Baseline (natural text)

- larger number of tokens are required
- sharper model growth as n-grams increase

System 2 (segmented text)

- model requires smaller corpus size
- model growth remains almost stable for higher n-grams

Effect in tokens and perplexity

System	Tokens	OOV	Perplexity
baseline	21042	1408	270
system 1	29808	757	69
system 2	29808	757	69

- Perplexity of language model decreased significantly
- Out of vocabulary rate reduced by more than a half compared to unsegmented model

Effect in MT performance

Metric	System	Avg
BLEU	baseline	15.6
	system 1	19.8
	system 2	20.9
METEOR	baseline	19.7
	system 1	21.1
	system 2	22.7
TER	baseline	74.2
	system 1	71.0
	system 2	72.7

- A slight improvement in BLEU and METEOR when using segmented corpus (System-2)
- TER (Translation Error Rate) – decreased using segmented corpus

Conclusion and Future plans

- English-Tigrinya **parallel corpus** was extracted and properly aligned
- Effect of morphological **segmentation** on English-Tigrinya statistical **machine translation** was investigated

Future plans:

- Improving **morphological segmentation** based on language model segmentation, semi-supervised or unsupervised approaches
- Investigating **SMT** on different levels of segmentation
- Investigating factored translation models

Ask yemane@jnlp.org
for detail.