Massively Multilingual Language Technologies

Building Bridges – Breaking Barriers
A Tribute to Alex Waibel, Professor, Pilot, Entrepreneur,...

Jaime Carbonell (www.cs.cmu.edu/~jgc)
Language Technologies Institute
Carnegie Mellon University

July 2016

The Many Faces of Alex Waibel

“Official” Alex

“Worried” Alex

“Happier” Alex

“Happiest” Alex

Werner Heisenberg v Alex Waibel

- **Heisenberg** Uncertainty Principle:
- It is impossible to measure location and momentum of an object precisely and simultaneously
- But... it can seem like we can measure them precisely if above the quantum level

- **Waibel** Uncertainty Principle:
- It is impossible to measure location and activity of Alex precisely and simultaneously
- But... it can seem like Alex is in multiple places doing multiple things at once

Bridging the Linguistic Divide

- **Until recently**: Focus was **Digital Divide**
  - The "digirati": internet connected, laptop,…
  - Smarphones → democratization in access
- **Currently**: The **Linguistic Divide** Rules
  - 6,900 languages: Google addresses 1%
  - Almost all information is in the 1%
  - How to democratize linguistic access?

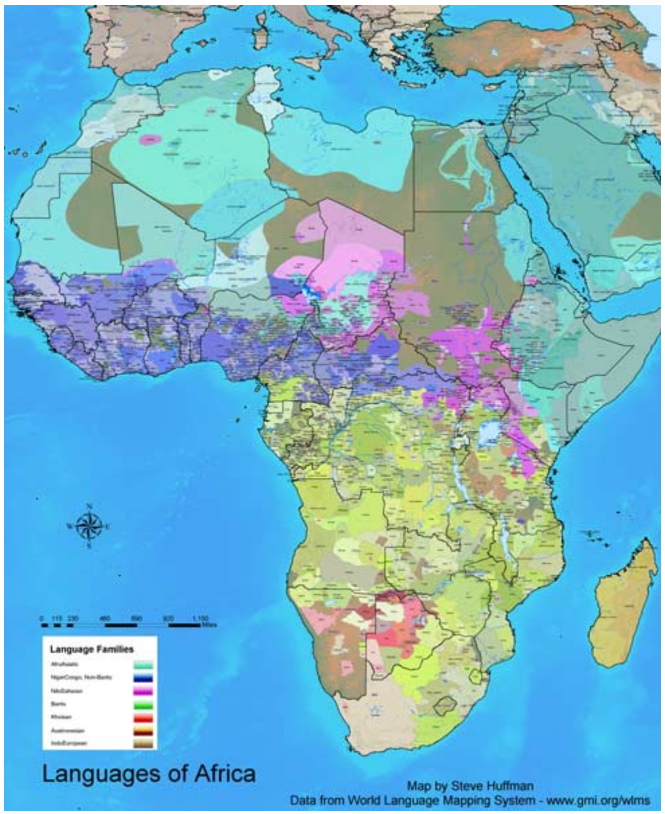
Multilingual Activities at CMU

- 1975: Speech research started at CMU (Harpy, Hearsay)
- 1986: Started Center for Machine Translation → LTI in 1996
- 1989: Knowledge-Based MT (domain specific, high accuracy)
- 1990: Multilingual speech (Sphinx, Janus)
- 1991: Example-Based MT
- 1992: Speech-to-speech MT (cStar)
- 1995: Statistical MT (Janus)
- 2000: MT for Low-Resource Languages
- 2006: Context-Based MT
- 2012: Linguistic-Core MT (for low resource languages)
- 45 Languages: English, Spanish, French, Japanese, German, Arabic, Korean, Chinese, Urdu/Hindi, Russian, Mapudungun, ...

Low Resource Languages

- 6,900 languages in 2012 – Ethnologue www.ethnologue.com/ethno_docs/distribution.asp?by=area
- Only 77 (1.2%) have over 10M speakers
  - 1st Chinese, 5th Arabic, 7th Bengali, 10th Javanese
- 3,000 have over 10,000 speakers each
- 3,000 may not survive past 2100
- 5X to 10X number of dialects (35 for Arabic)
- \# of L's in some interesting countries:
  - Afghanistan: 52, Pakistan: 77, India 400
  - North Korea: 1, Indonesia 700

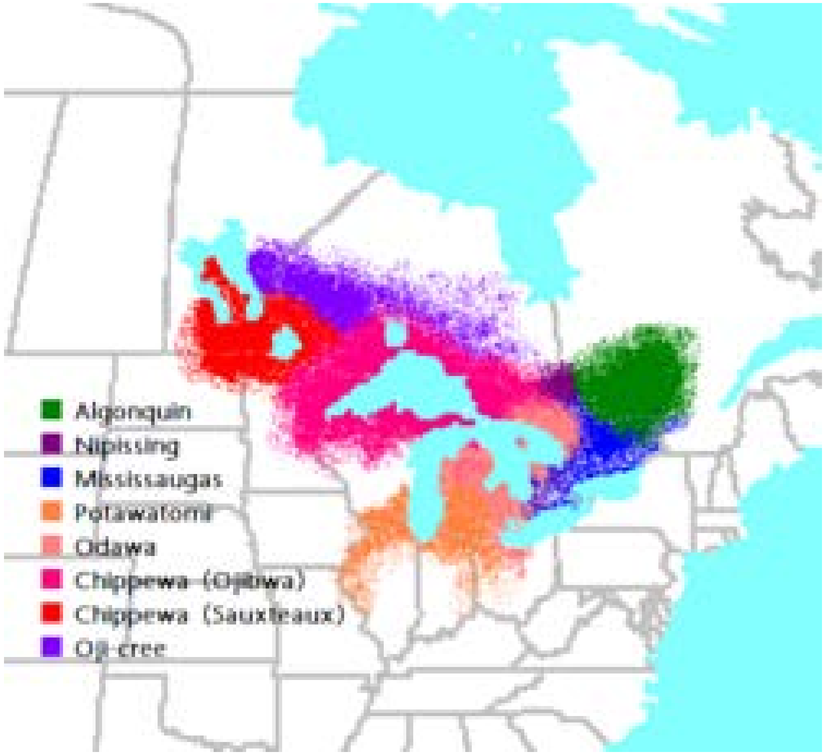
Some Linguistics Maps



Some (very) LD Languages in the US

Anishinaabe (Ojibwe, Potawatame, Odawa)

Great Lakes



Challenges for General MT

- Ambiguity Resolution
  - Lexical, phrasal, structural
- Structural divergence
  - Reordering, vanishing/appearing words, ...
- Inflectional morphology
  - Spanish 40+ verb conjugations, Arabic has more.
  - Mapudungun, Anupiac, ... → agglomerative
- Training Data
  - Bilingual corpora, aligned corpora, annotated corpora, bilingual dictionaries
- Human informants
  - Trained linguists, lexicographers, translators
  - Untrained bilingual speakers (e.g. crowd sourcing)
- Evaluation
  - Automated (BLEU, METEOR, TER) vs HTER vs ...

Context Needed to Resolve Ambiguity

Example: English → Japanese

- Power **line** – densen (電線)
- Subway **line** – chikatetsu (地下鉄)
- (Be) on **line** – onrain (オンライン)
- (Be) on the **line** – denwachuu (電話中)
- **Line** up – narabu (並ぶ)
- **Line** one's pockets – kanemochi ni naru (金持ちになる)
- **Line** one's jacket – uwagi o nijuu ni suru (上着を二重にする)
- Actor's **line** – serifu (セリフ)
- Get a **line** on – joho o eru (情報を得る)

Sometimes local context suffices (as above) → n-grams help

. . . but sometimes not

CONTEXT: More is Better

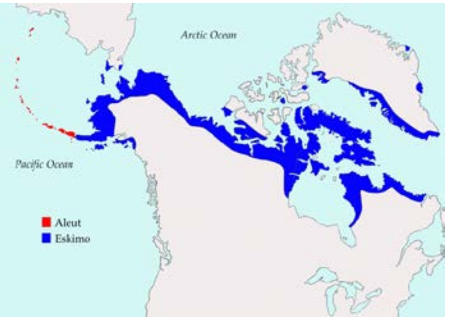
- **Examples requiring longer-range context:**
  - "The **line** for the new play **extended for 3 blocks**."
  - "The **line** for the new play was changed by the **scriptwriter**."
  - "The **line** for the new play got **tangled with the other props**."
  - "The **line** for the **new play** better protected the **quarterback**."
- **Challenges:**
  - Short n-grams (3-4 words) insufficient
  - Requires more general syntax & semantics

Additional Challenges for LD MT

- Morpho-syntactics is plentiful
  - Beyond inflection: verb-incorporation, agglomeration, ...
- Data is scarce
  - Insignificant bilingual or annotated data
- Fluent computational linguists are scarce
  - Field linguists know LD languages best
- Standardization is scarce
  - Orthographic, dialectal, rapid evolution, ...

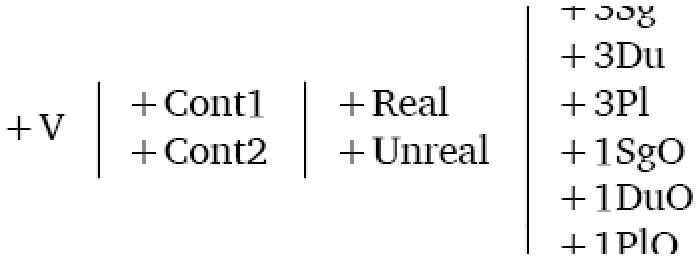
Morpho-Syntactics & Multi-Morphemics

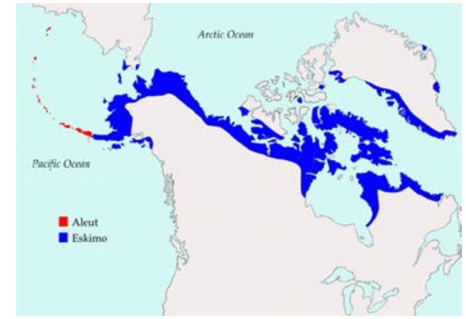
- Iñupiaq (North Slope Alaska, Lori Levin)
  - Tauqsiġñiaġviŋmuŋniaŋitchugut.
  - ‘We won’t go to the store.’

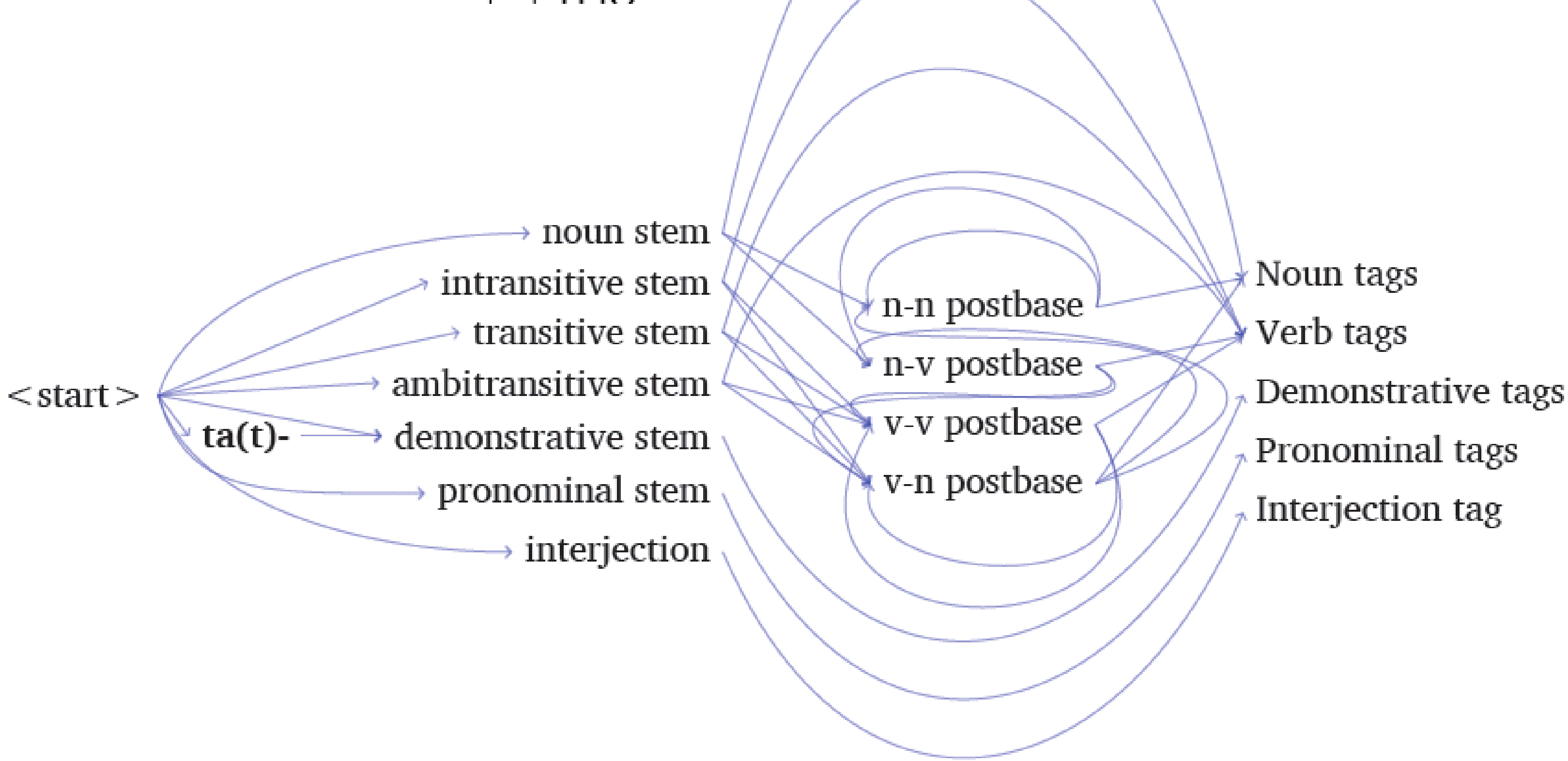


- Kalaallisut (Greenlandic, Per Langaard)
  - Pittsburghimukarthussaqarnavianngilaq
  - Pittsburgh+PROP+Trim+SG+kar+tuq+ssaq+qar+naviar+nngit+v+IND+3SG
  - "It is not likely that anyone is going to Pittsburgh"

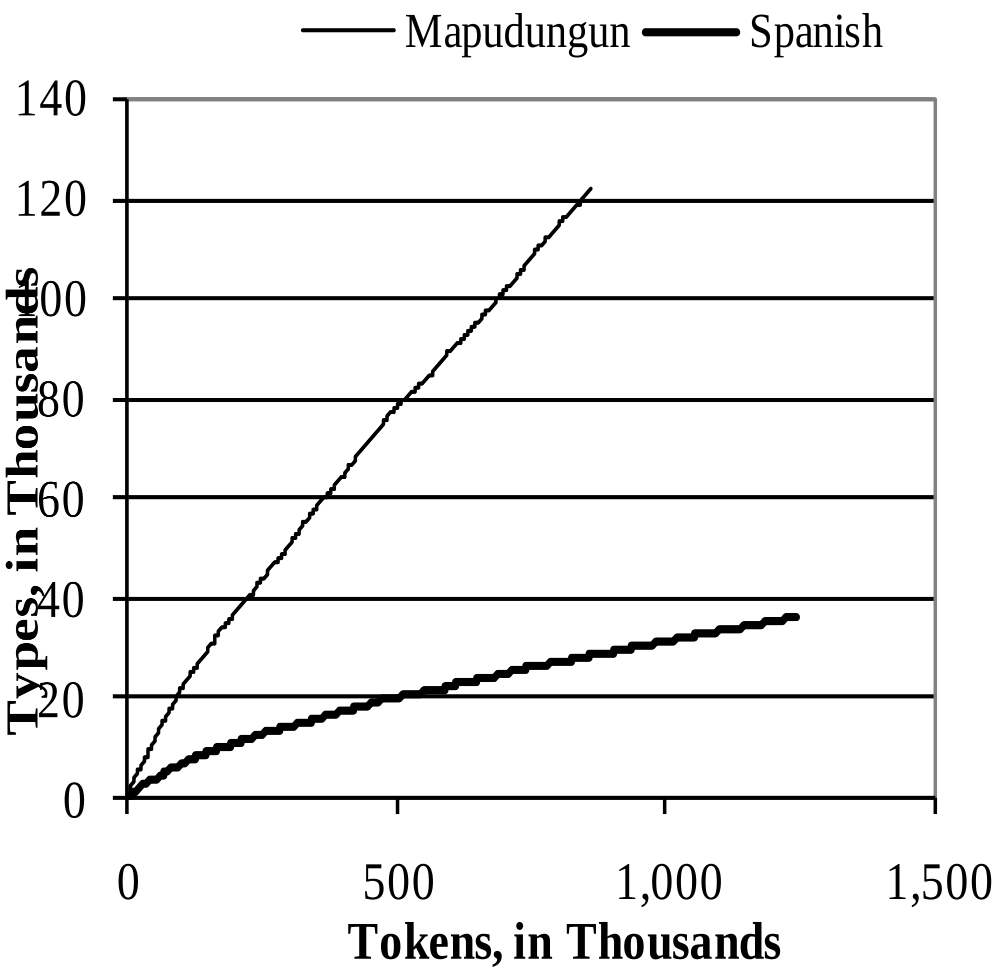
Morphotactics in Iñupiaq





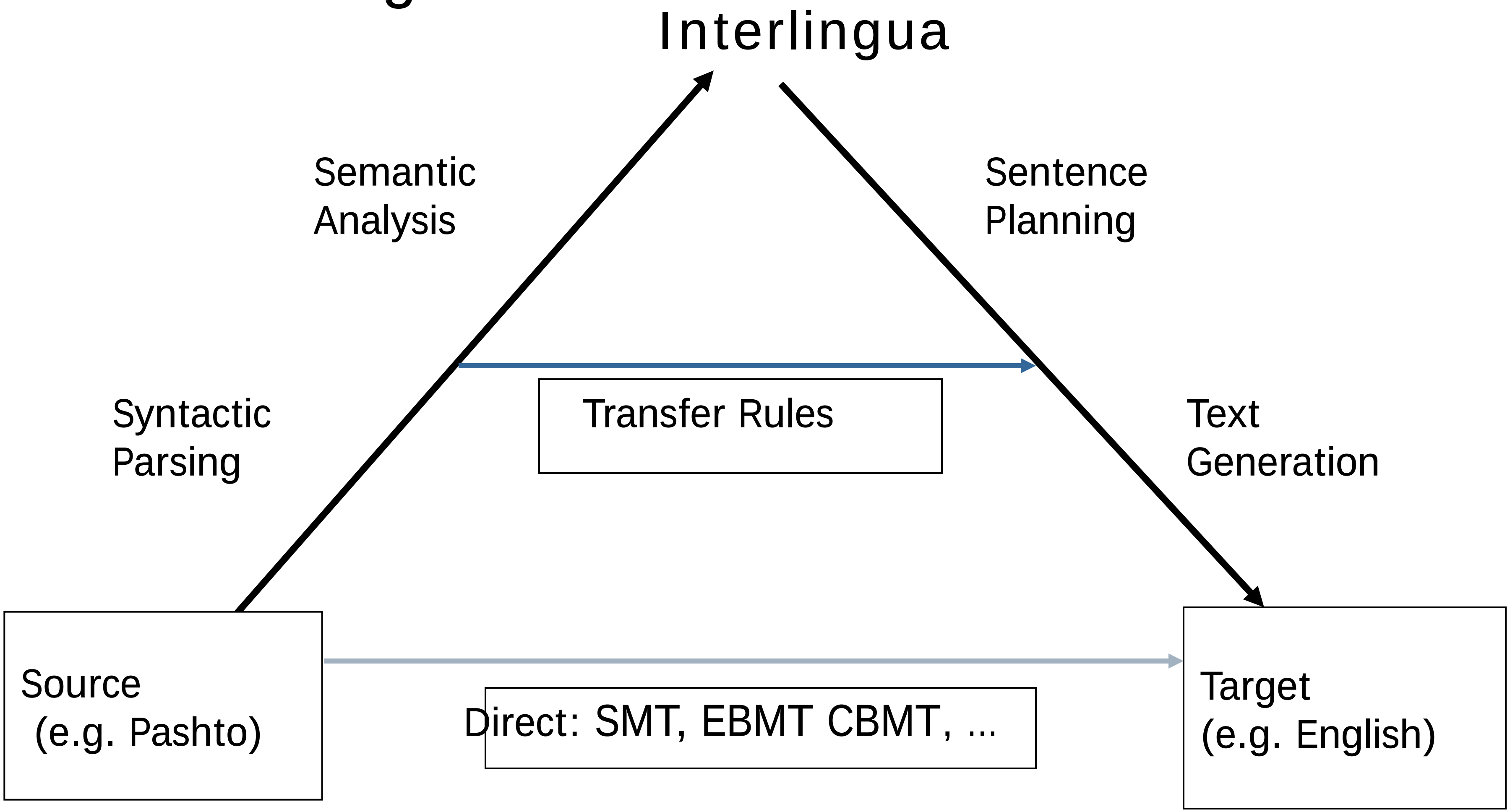


Type-Token Curve for Mapudungun

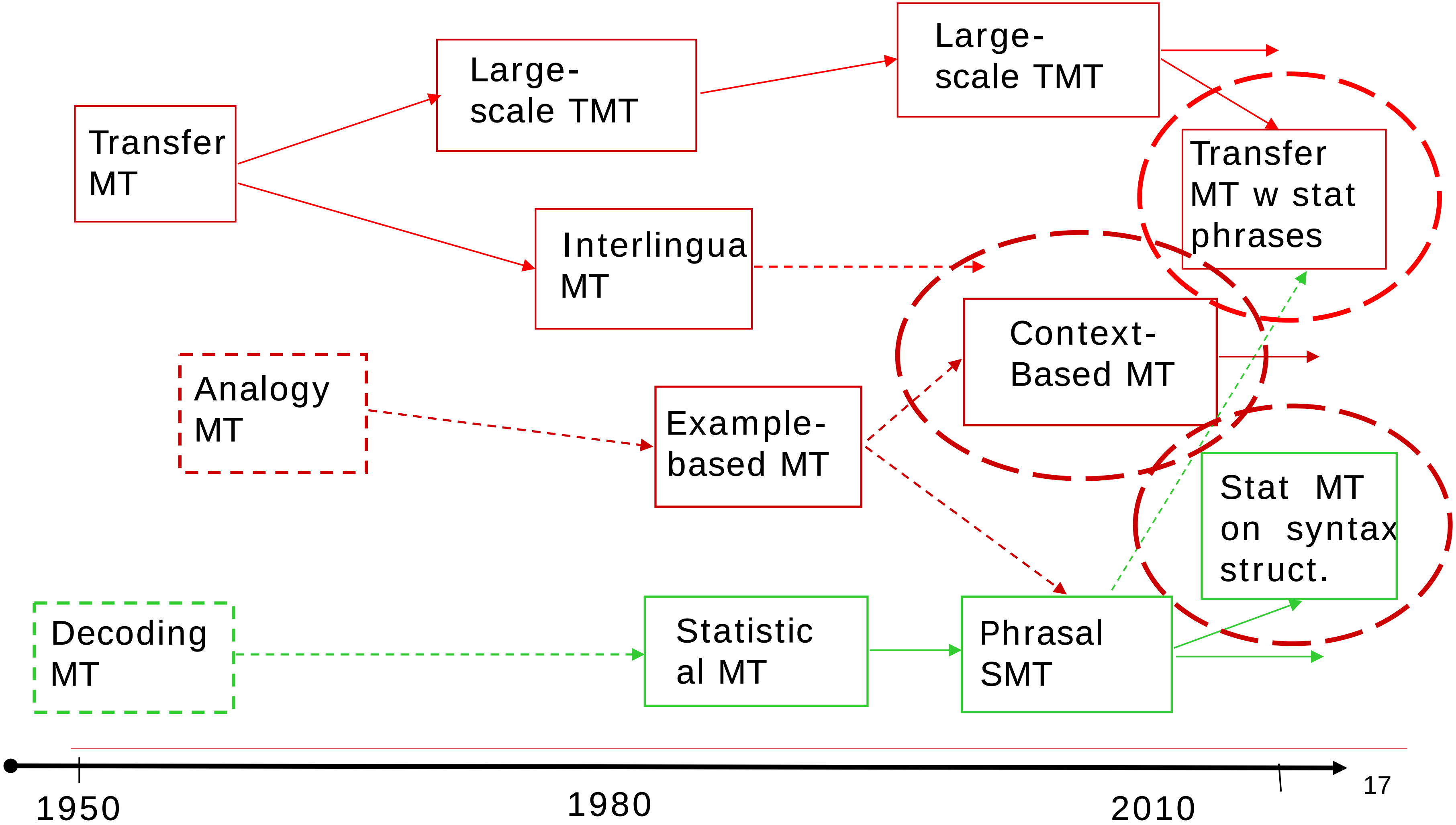


- 400,000+ speakers
- Mostly bilingual
- Mostly in Chile
    - Pewenche
    - Lafkenche
    - Nguluche
    - Huilliche

Paradigms for Machine Translation



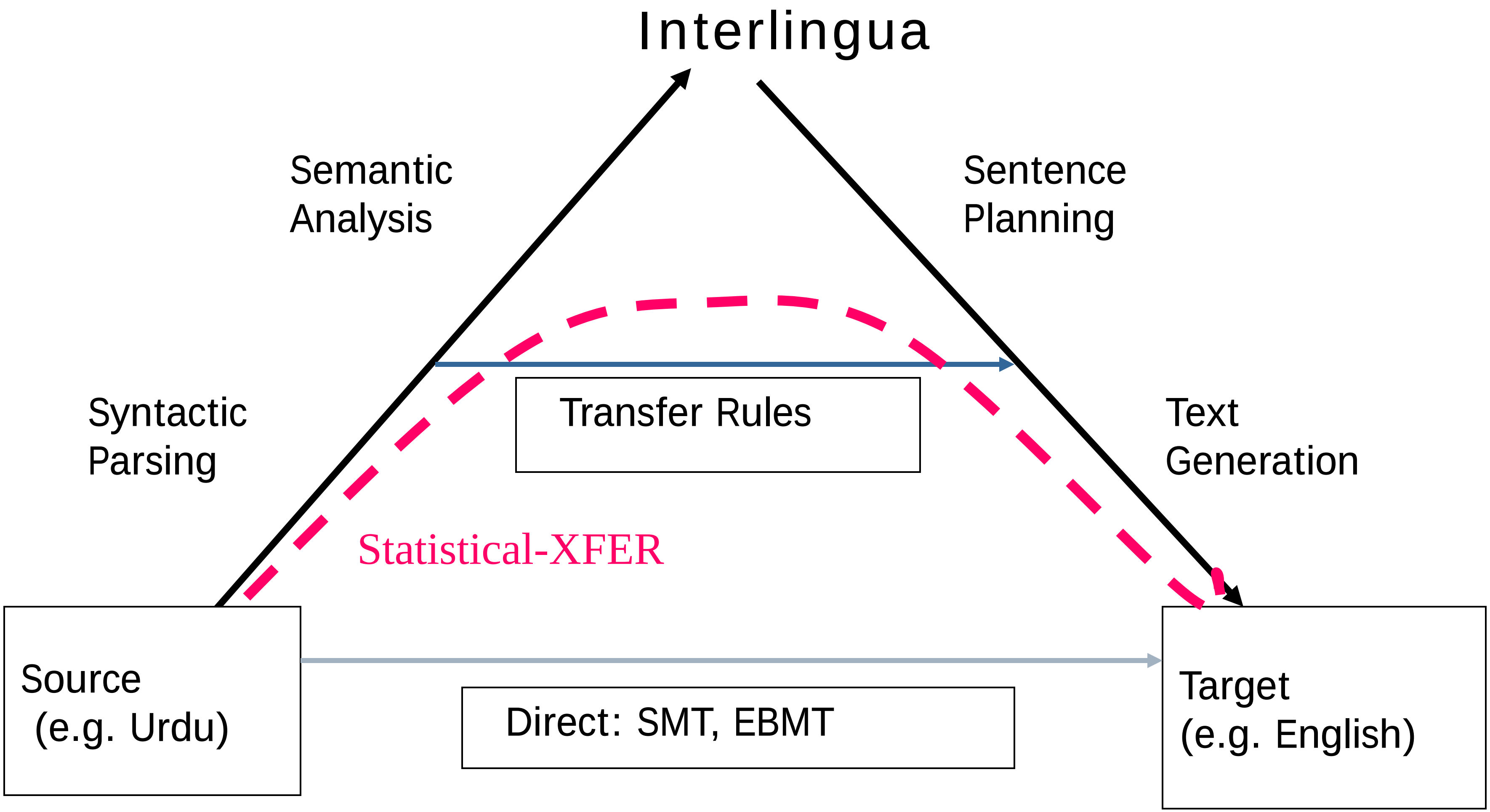
Evolutionary Tree of MT Paradigms



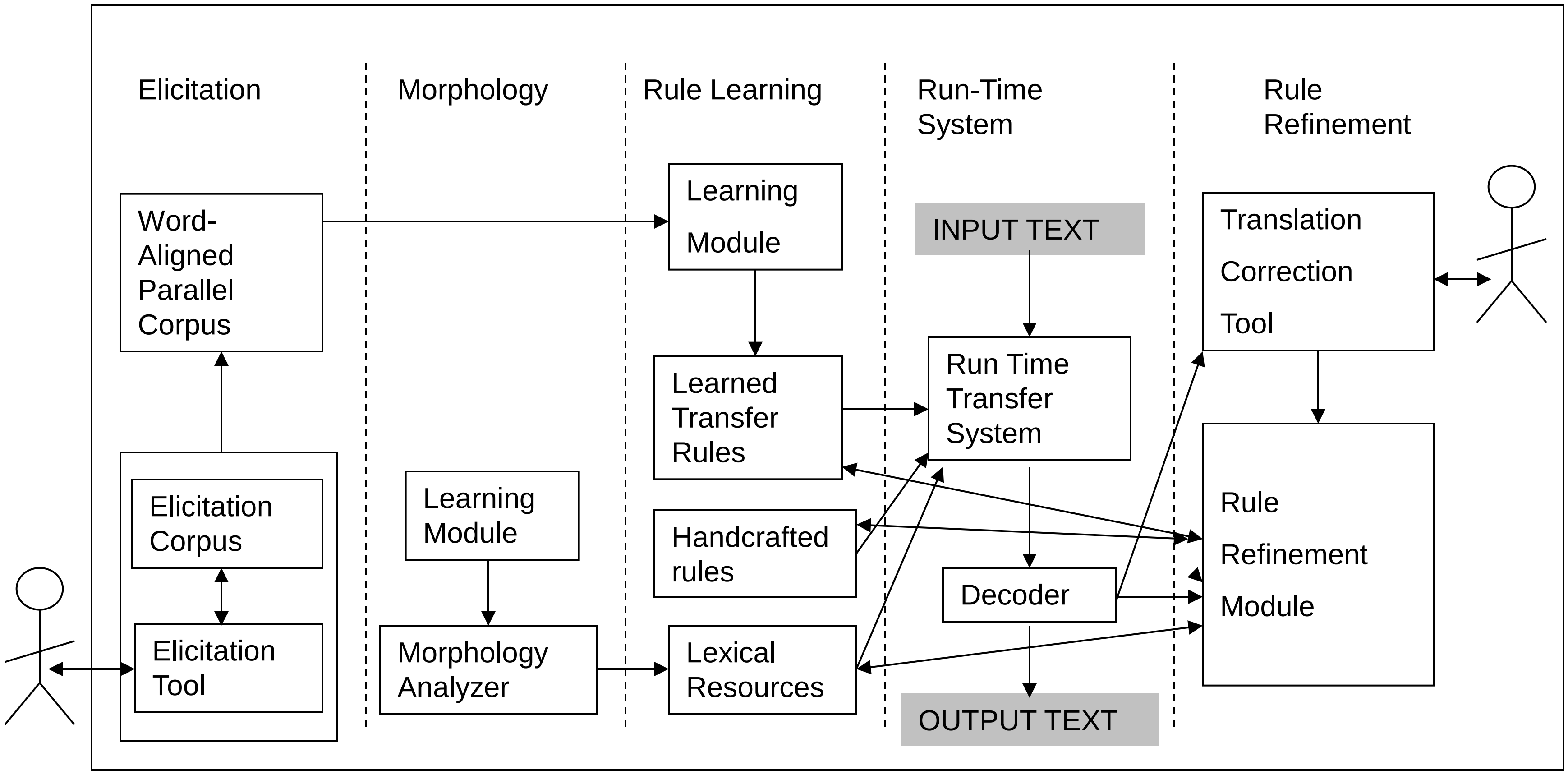
Stat-Transfer (STMT): List of Ingredients

- **Framework:** Statistical search-based approach with syntactic translation transfer rules that can be acquired from data but also developed and extended by experts
- **SMT-Phrasal Base:** Automatic Word and Phrase translation lexicon acquisition from parallel data
- **Transfer-rule Learning**: apply ML-based methods to automatically acquire syntactic transfer rules for translation between the two languages
- **Elicitation**: use bilingual native informants to produce a small high-quality word-aligned bilingual corpus of translated phrases and sentences
- **Rule Refinement**: refine the acquired rules via a process of interaction with bilingual informants
- **XFER + Decoder**:
  - XFER engine produces a lattice of possible transferred structures at all levels
  - Decoder searches and selects the best scoring combination

Stat-Transfer (ST) MT Approach



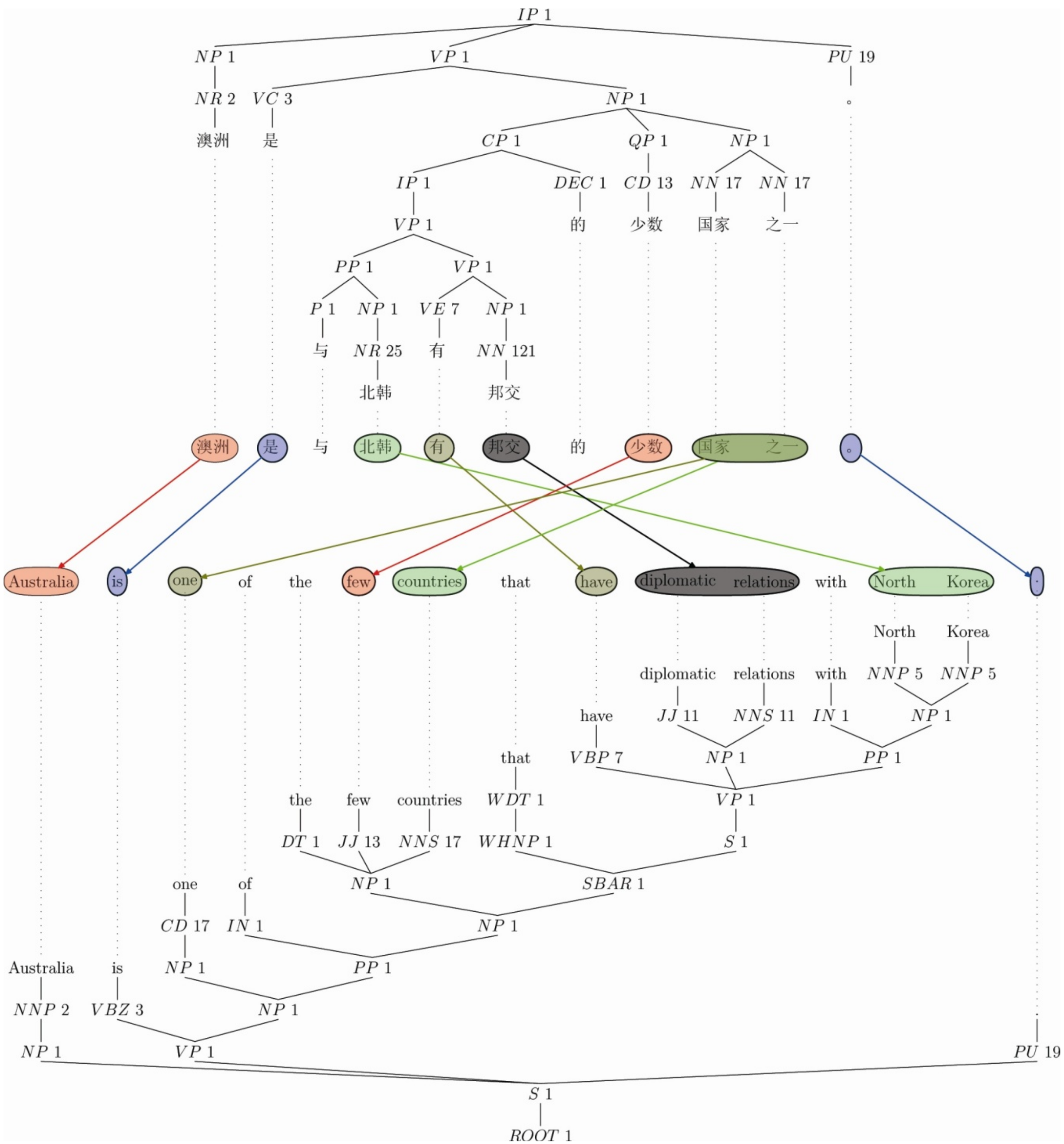
Avenue/Letras STMT Architecture



Syntax-driven Acquisition Process

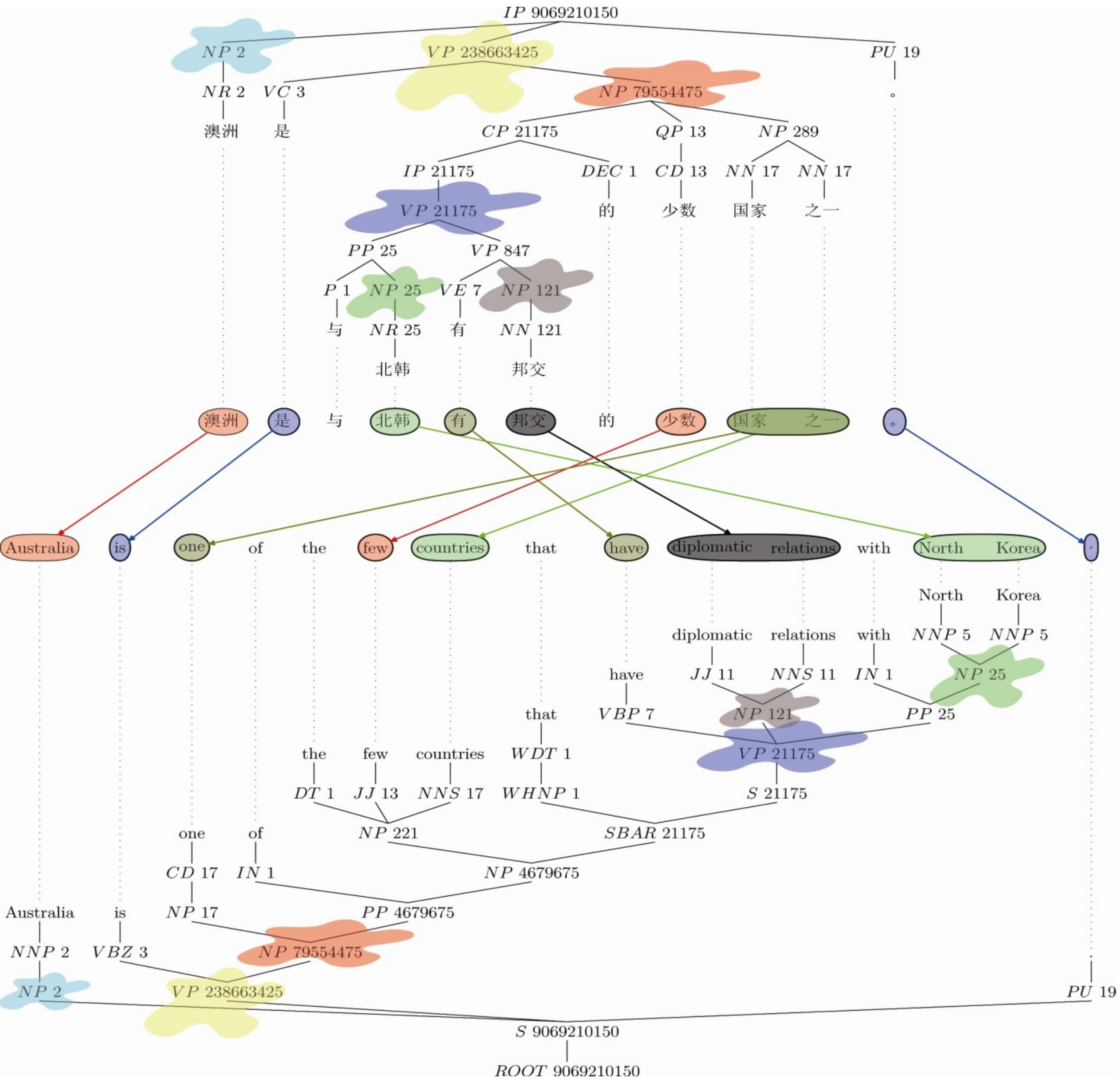
Automatic Process for Extracting Syntax-driven Rules and Lexicons from sentence-parallel data:

1. **Word-align** the parallel corpus (GIZA++)
2. **Parse the sentences independently** for both languages
3. **Tree-to-tree Constituent Alignment**:
   a) **Run our new Constituent Aligner** over the parsed sentence pairs
   b) **Enhance alignments** with additional Constituent Projections
4. **Extract all aligned constituents** from the parallel trees
5. **Extract all derived synchronous transfer rules** from the constituent-aligned parallel trees
6. **Construct a "data-base"** of all extracted parallel constituents and synchronous rules **with their frequencies** and model them statistically (assign them **relative-likelihood probabilities**)



PFA Node Alignment Algorithm Example

- Any constituent or sub-constituent is a candidate for alignment
- Triggered by word/phrase alignments
- Tree Structures can be highly divergent



PFA Node Alignment Algorithm Example

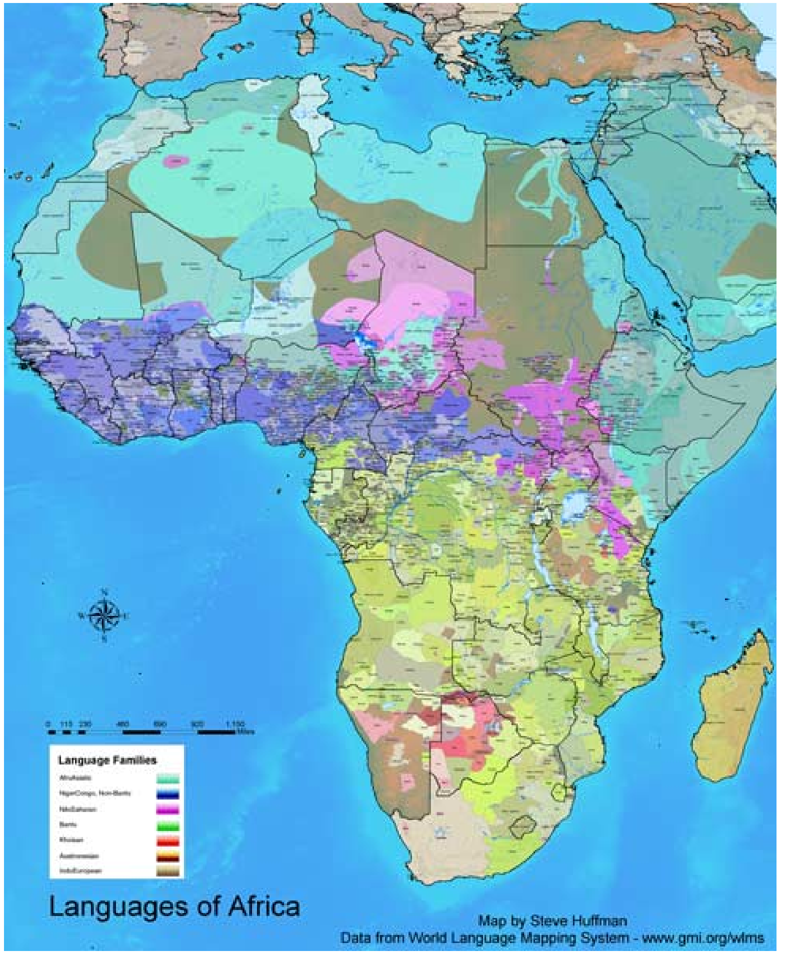
- Tree-tree aligner enforces equivalence constraints and optimizes over terminal alignment scores (words/phrases)

- Resulting aligned nodes are highlighted in figure

- Transfer rules are partially lexicalized and read off tree.

The Setting

- MURI Languages
  - Kinyarwanda
    - Bantu (7.5M speakers)
  - Malagasy
    - Malayo-Polynesian (14.5M)
  - Swahili
    - Bantu (5M native, 150M 2nd/3rd)



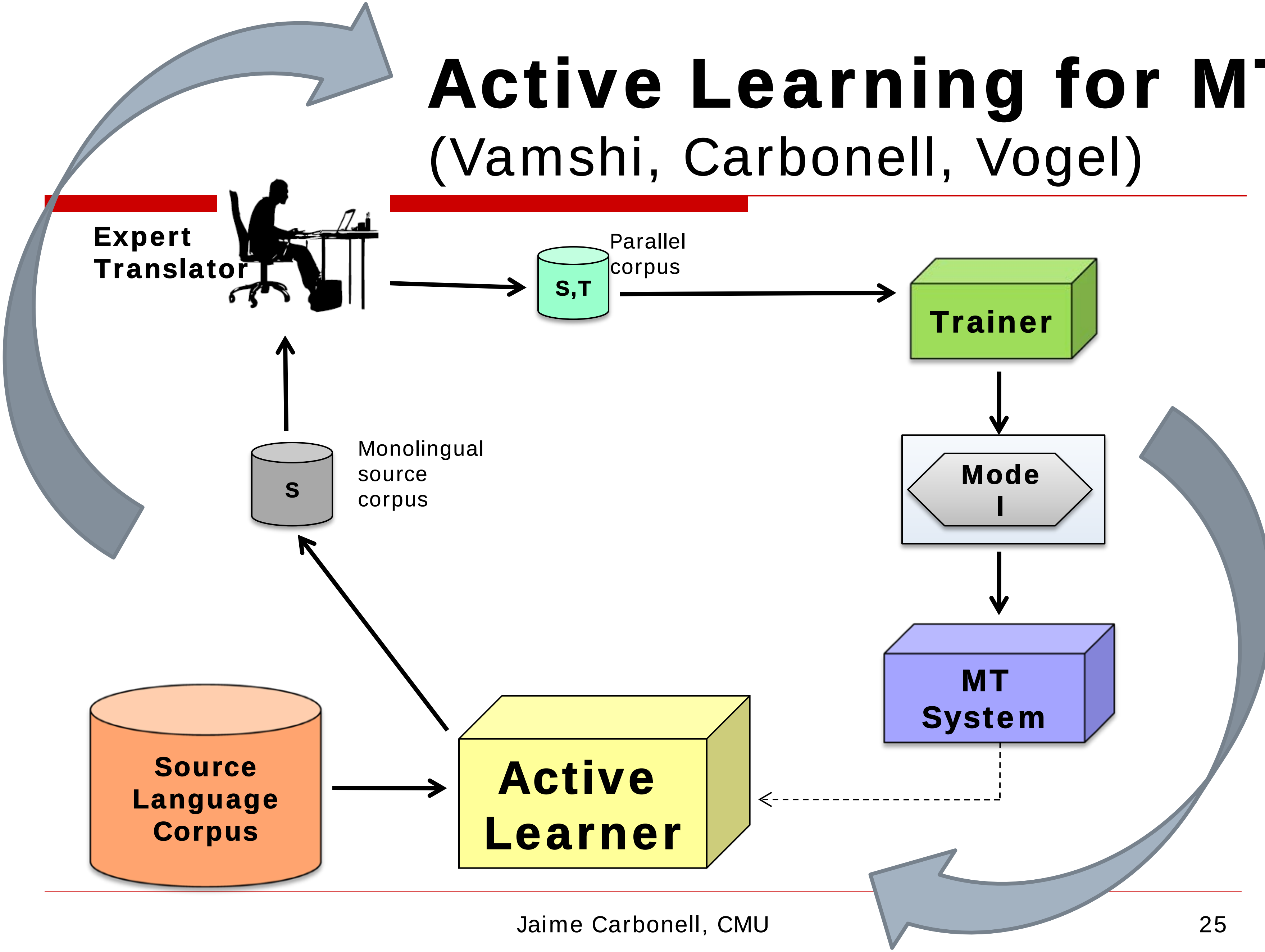
Swahili

Anamwona

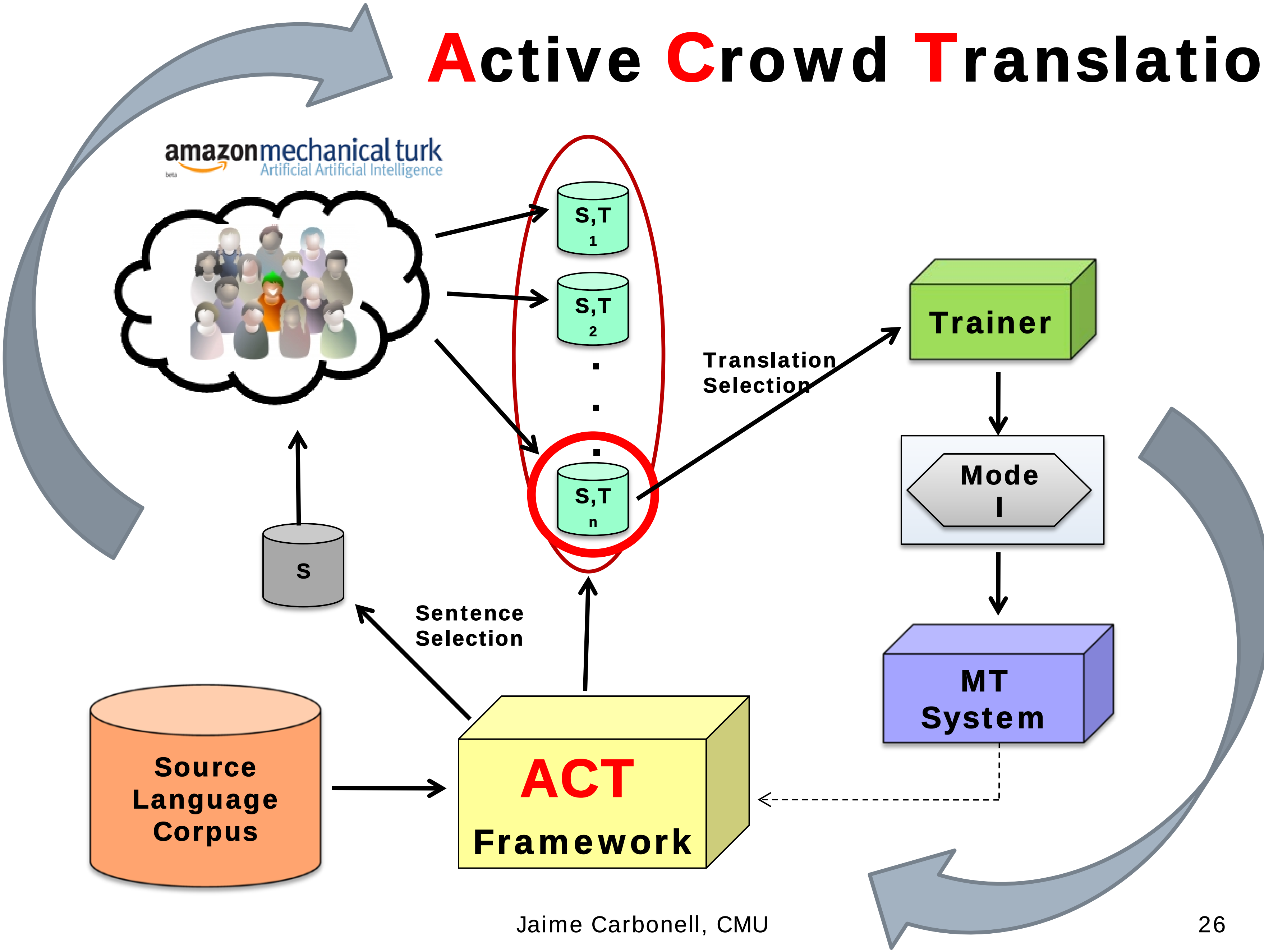
"he is seeing him/her"

→Morpho-syntactics

Active Learning for MT
(Vamshi, Carbonell, Vogel)



Active Crowd Translation



Active Learning Strategy: Diminishing Density Weighted Diversity Sampling

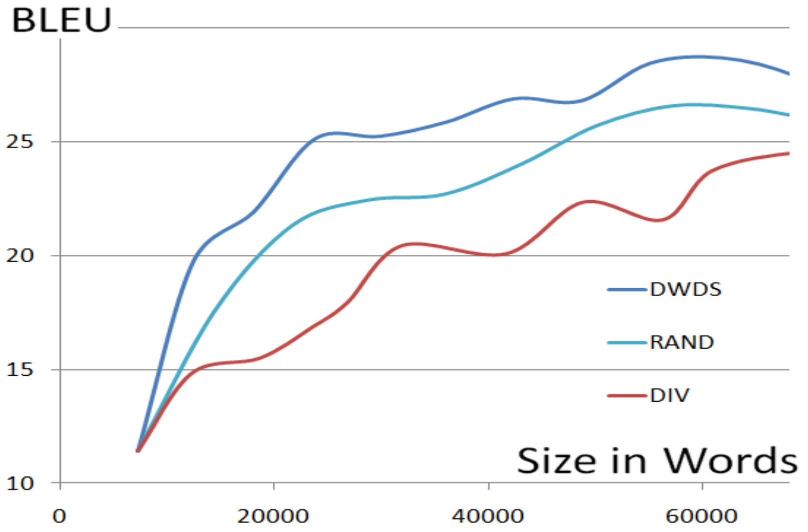
$$density(S) = \frac{\sum_{x \in Phrases(s)} P(x/UL) * e^{\wedge}-[\lambda * count(x/L)]}{|Phrases(s)|}$$

$$diversity(S) = \frac{\sum_{x \in Phrases(s)} \alpha * count(x)}{|Phrases(s)|}$$

$$\alpha = 0 if x \in L$$

$$\alpha = 1 if x \notin L$$

$$Score(S) = \frac{(1+\beta^2) density(S) * diversity(S)}{\beta^2 density(S) + diversity(S)}$$



Experiments:
Language Pair: Spanish-English
Batch Size: 1000 sentences each
Translation: Moses Phrase SMT
Development Set: 343 sens
Test Set: 506 sens

Graph:
X: Performance (BLEU)
Y: Data (Thousand words)

Translation Selection from Mechanical Turk

- Translator Reliability

$$rel(W_k) = \frac{\sum_{t_j \in T_k} \sum_{n_i \in U} \alpha}{\|T_k\|}$$

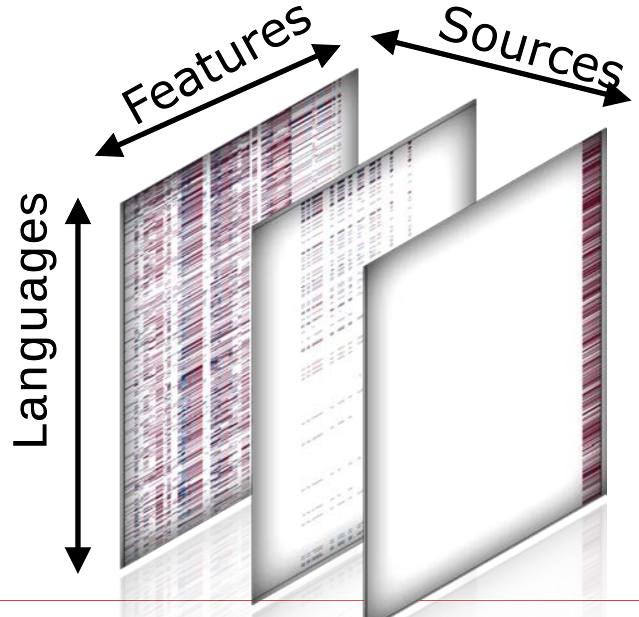
$$\alpha = \begin{cases} 1 & t_{kj} \equiv t_{nj} \\ 0 & \end{cases}$$

- Translation Selection:

	Seed	Iterations	
System	0	1	2
crowd pick-rand	10.64	18.64	21.07
crowd translation-agreement	10.64	21.81	24.67
crowd translator-agreement	10.64	22.78	24.94
expert translations	10.64	22.34	25.75
crowd all-three	10.64	23.68	26.01

ARIEL: Universal Typological Compendium

- **Curation**: ~2000 typological features across ~2200 languages
- **Distance functions**: Phylogenetic, Geopolitical, Typological



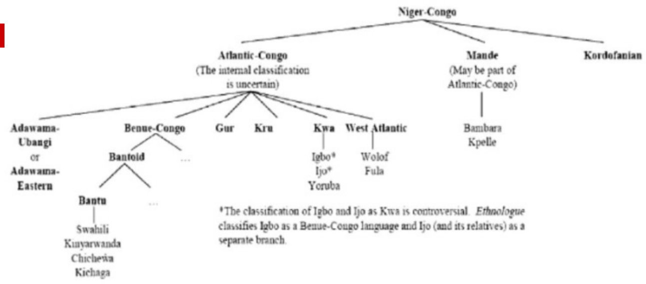
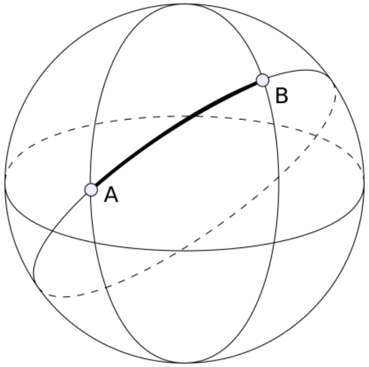
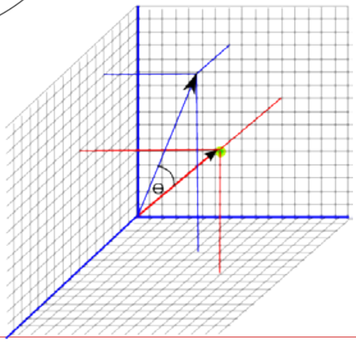
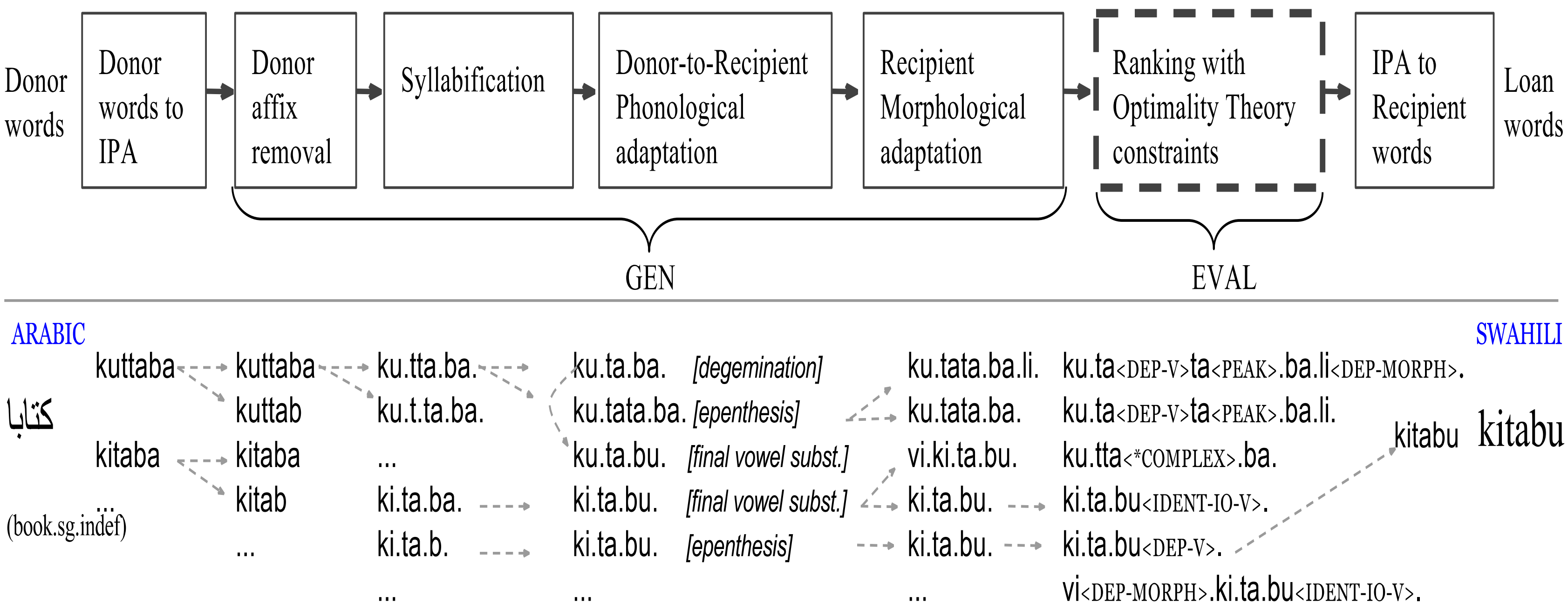


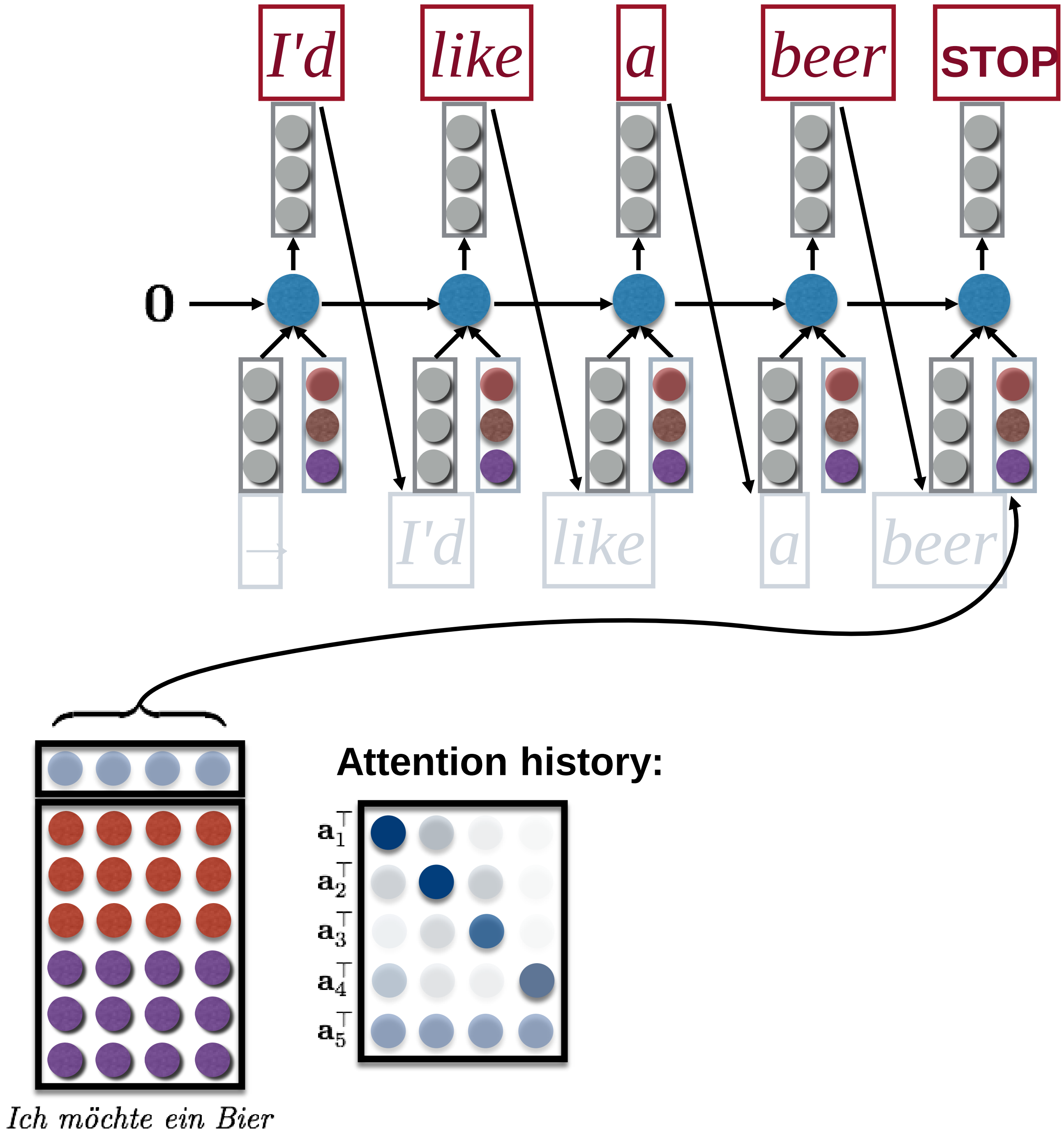
Figure 6.1. *The Niger-Congo language family.*





Lexical Transfer Example
(Arabic → Swhahili)




I'd
like
a
beer
STOP
0
I'd
like
a
beer
Attention history:
$\mathbf{a}_1^\top$
$\mathbf{a}_2^\top$
$\mathbf{a}_3^\top$
$\mathbf{a}_4^\top$
$\mathbf{a}_5^\top$
Ich möchte ein Bier

Concluding Remarks

- Massively multilingual research (MT, speech, dialog)
  - Of crucial importance for humanity
  - Waibel has been at the very core
- Research Directions
  - Combining linguistics and Statistics
  - Paradigms for cross-language scalability
  - Transfer learning and proactive learning
  - Applications: disaster relief, education, eCom, ...

THANK YOU!