Haitian Creole

Developing MT for a Low Data Language

William Lewis

Microsoft Research

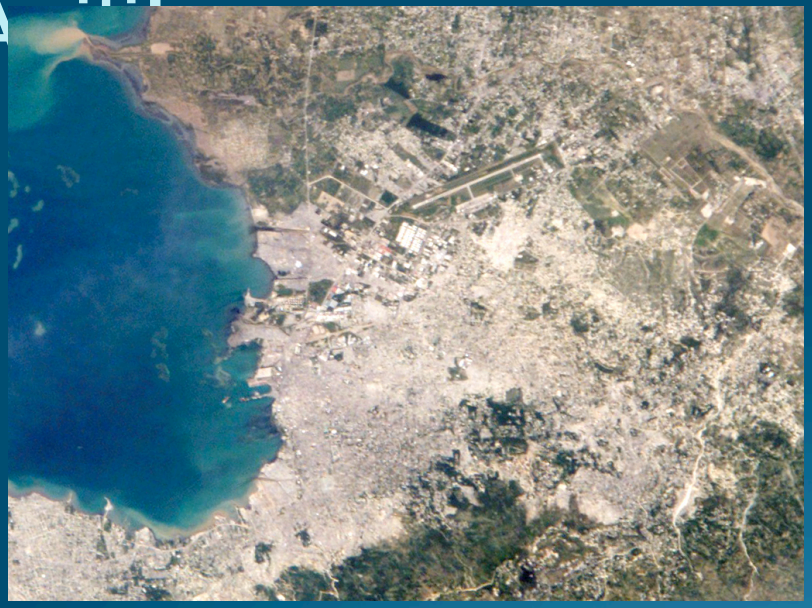
Credits

- Carnegie Mellon University
- Butler Hill Group
- Mission 4636/Crowdflower
- Ushahidi
- Moravia Worldwide
- Welocalize
- Rosetta Foundation
- Eriksen Translations, Inc.
- The Bing Team
- All members of the Microsoft Translator team who put in many sleepless nights on this project.

Haitian Creole

- One of two official languages in Haiti
- A creole that evolved from French, Spanish, and several African languages (large % French-like)
- Spoken natively by most of Haiti's 8M people
- Recent as a written language (first literature dates to late 18th century), growing literature base
- Semi-literate population, with preference to French (until recently)
- Somewhat inconsistent orthography
- Limited (but growing) Web presence

Tranbleman tè nan Pòtoprens, kapital A[illegible]



Pòtoprens te catastrophically afekte 12 janvye 2010 tranbleman tè a.

- The earthquake of January 12th, 2010 a significant humanitarian crisis.
- Aid agencies, foreign governments, a variety of NGOs, all responded en masse

Moun ap fouye pami debri yon bilding ki kraze nan tranblemann' tè 12 Janvye a.

- Need for translated materials critical, especially those related to medicine and the relief effort.
- Mission 4636 text messages from the field (up to 5K/hour at peak) require rapid translation

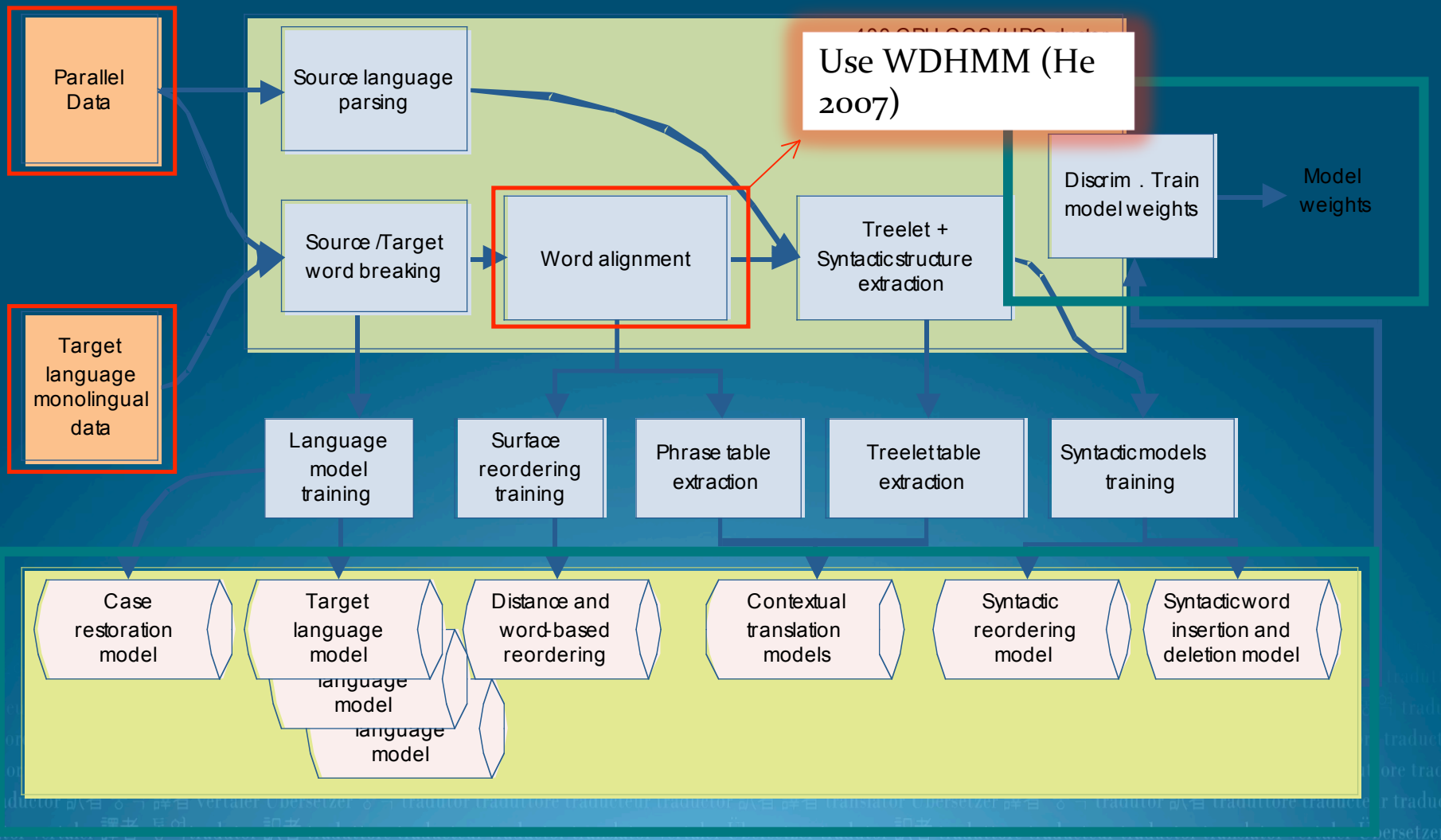
The E-mail

- At 10:30 a.m. on Tuesday, January 19th our team received an e-mail from a Microsoft employee in the field:
  - Do we have a translator for Haitian Creole?
  - If not, could we make one?
- A little soul searching:
  - No one on our team knew anything about Creole
    - No native speakers
    - No linguistic background on the language
    - No idea about grammatical structure
  - No idea about encoding or orthography
  - No knowledge about registers or the degree of literacy
  - No parallel or monolingual training data of any kind (nor readily available documents we could start with)
  - In effect, we were starting at **Zero**
- So what else could we do but say "YES!"

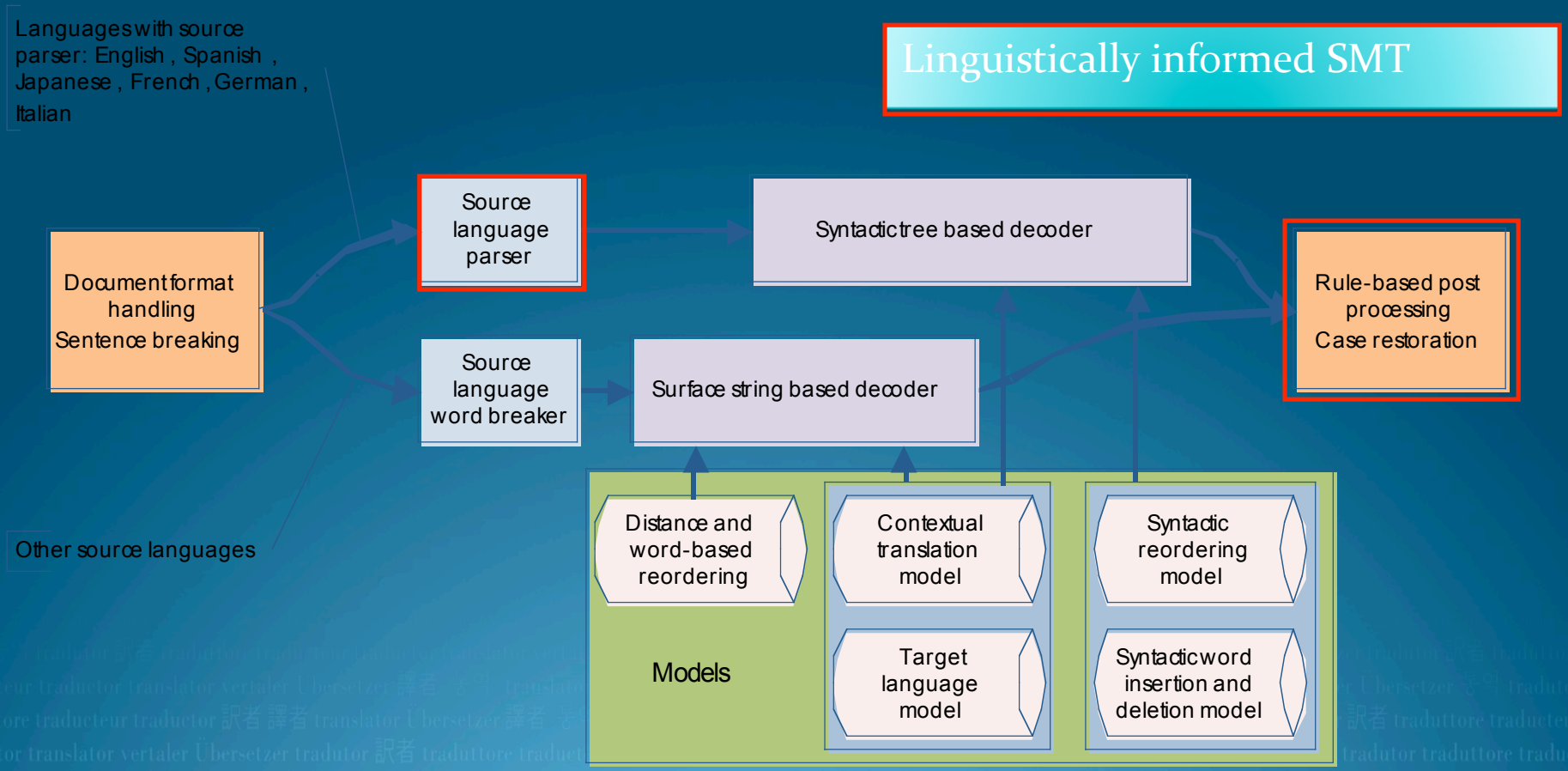
The Plan

- Identify as much parallel data as we can find; start with
  - Bible
  - Data from Carnegie Mellon University (CMU)
  - Haitisurf.com
  - Official government documents, including constitution
  - Data identified by CrisisCommons
  - Parallel sentences from Creole-English Wiki pages
- Rally team to help process the data (and everything else!)
- Find linguistic experts in Creole to advise and help
- Find native speakers to review output and translate content
- Engage the relief community involved in the Haiti effort

Training


Parallel Data
Target language monolingual data
Source language parsing
Source /Target word breaking
Word alignment
Use WDHMM (He 2007)
Treelet + Syntactic structure extraction
Discrim . Train model weights
Model weights
Language model training
Surface reordering training
Phrase table extraction
Treelet table extraction
Syntactic models training
Case restoration model
Target language model
Distance and word-based reordering
Contextual translation models
Syntactic reordering model
Syntactic word insertion and deletion model

Microsoft's Statistical MT Engine



Previous work on low-data MT

Low data MT not without precedent:

- DARPA sponsored Surprise Language Exercise (SLE)
  - One month to collect data, create resources (Oard 2003)
  - Initial test case Cebuano (Strassel et al 2003)
  - One month competition on Hindi (multiple teams)
- Oard and Och 2003 relate effort to rapidly develop MT over data collected in SLE
  - Noted that MT could be developed "in days"

- Haitian specific work:
  - DIPLOMAT project (Frederking et al 1997)
    - Speech-to-Speech translation system
    - Shelved, but data housed at CMU

Challenges presented by Creole

- Low Data
- Creole "young" as a written language, inconsistent orthography (Allen 1998)
- Two "registers" in written form:
  - High register: full forms for pronouns and function words
  - Low register: contracted forms, but inconsistent

Pronoun	Gloss	Appears as
mwen	I, me, mine	m, 'm, m'
nou	you (pl), us	n, 'n, n'
ou	you	w, w'
li	he, she, it	l, l', 'l

Challenges presented by Creole

- Low Register also has large number of reduced forms:

Abbreviated Form	Full Form
s'on	se yon
avèn	avèk nou
relem	rele mwen
wap	ou ap
map	mwen ap
zanmim	zanmi mwen
lavel	lave li
…	…

- Has three accented characters, è, ò, à
  - Accents inconsistently used, especially in SMS, e.g., mesi vs. mèsi, le vs. lè
- Inconsistent compounding: tranblemantè', tranbleman tè, tranbleman de tè' -- "earthquake"

Processing and Filtering Data

- Focused on reducing data sparseness
- Forced separation of data sets between English-Creole (EC) vs. Creole-English (CE)
- For CE:
  - Normalized out all accented forms
  - Likewise, normalized contracted and reduced forms to full forms
  - Did the same at run time
- For EC:
  - Significant normalization not possible w/o introducing noise
  - Some post-processing repairs possible (i.e., in our rule-based post-processing component)

The Timeline

- Tues., January 19th, 10:30 a.m.: Email received
- Tues. afternoon: decision made, team rallied: developers, testers, computational linguists engaged
- Tues. afternoon: initial design on dev lead's whiteboard
- Wed. morning: division of labor established, small team dedicated to data collection and processing
- Wed. afternoon: first data sources processed (e.g., CMU, Bible, etc.)
- Wed. afternoon: clear division in CE and EC data
- Wed. evening: started assembling first configs for training systems
- Thurs., 4:00 a.m.: first training started
- Thurs., 10:45 a.m.: bug found in CMU data, fixed and reported to CMU (misalignment, reversed languages)
- Thurs., 2:15 p.m.: first successful build, Creole-English, BLEU score of 22.94 on held-out CMU data!
- Fri. morning: first Creole linguists, translators engaged
- Fri. & Sat.: continued data procurement, training, consulting with linguists and native speakers

Chasing the Chickens

(rolling it out)

- Saturday, 4:49pm – language models done, check in & start data push
- 5:00pm – leaf machines not translating Creole
- 5:33pm – processing out of sync, restart everything. Translations again!
- 5:53pm – deploy 3rd build to test environment
- 6:12pm – find 100K more parallel sentences, should we take them? YES!
- 6:14pm – in a sign of eternal optimism, take one prod offline
- 6:52pm – test 3rd rollout done, start testing everything
- 7:21pm – something's wrong, it's *really* slow
- 8:11pm – pour through ~1GB of logs trying to figure out what's wrong
- 8:49pm – find golden sentence mismatch (sanity check)
- 9:09pm – fix golden sentences
- 10:40pm – 4th build done
- 10:42pm – deploy 4th build to test
- 11:38pm – deploy done. Start testing it

Chasing the Chickens (con't)

- Sunday, 12:05am – "The united states believe this ideal right of chickens do the birth…"
- 12:05am – problem parsing smart quotes
- 1:06am – hot fix smart quotes for chickens
- 1:20am – chickens are gone
- 1:36am – Ship it! Begin rollout to prod
- 2:09am – rollout done. Start testing and warm-up
- 2:48am – load tests look good
- 3:30am – rollout done
- 3:31am – load test and warm-up
- 4:00am – load tests look good
- 4:01am, January 24th (Sunday) – prod live. We're done!
- Start to finish (from e-mail to ship): 4 Days, 17 Hours, and 31 Minutes

Where we are and Where we're going

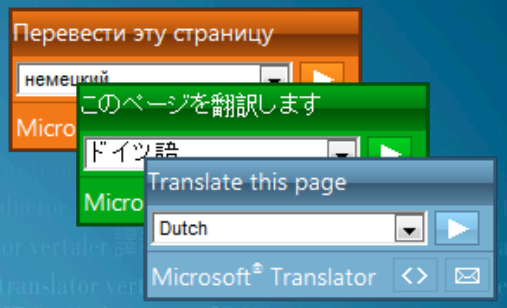
- Current BLEU:
  - CE: 29.89, EC: 18.30
  - Eval data:
    - 550 segments held-out CMU data, plus
    - 36 SMS messages (more in soon to be updated version)
  - Training data currently >200K segments (initial system: ~80K)
- Continued improvements through additional data
- Tapping English-French vocab, and English-French / English-Creole ASR dictionaries for OOV reduction (CE only)
- Continued Engagement with Crowdflower/Mission 4636
  - Translating and repairing SMS content
  - Initial supply of 1,000 SMS messages given back to Mission 4636
  - Once anonymized, all data (~5,000 SMS messages) will be provided back to Mission 4636 and the greater community (through CMU, LDC, TAUS TDA and the Rosetta Foundation)

Mission 4636 Messages

Mwen rele FIRST LAST mwen	My name is FIRST LAST. I
se yon bòs mason	work in construction,
kay mwen kraze mwen gen	and I have four children.
kat pitit numero mwen	My number is 99999999.
se 99999999	
Ki sa pou nou f? ak timoun	What can we do with the
yo kos?nan lekol la e pui	children regarding school
kile moun duval nan croi	and when will the people
des bouket ap jwen manje	of duval in croix des
pou met nan vant yo	bouquets get food to put
	in their bellies?
Voye kÄk konsÄy pou	Send me some advice.

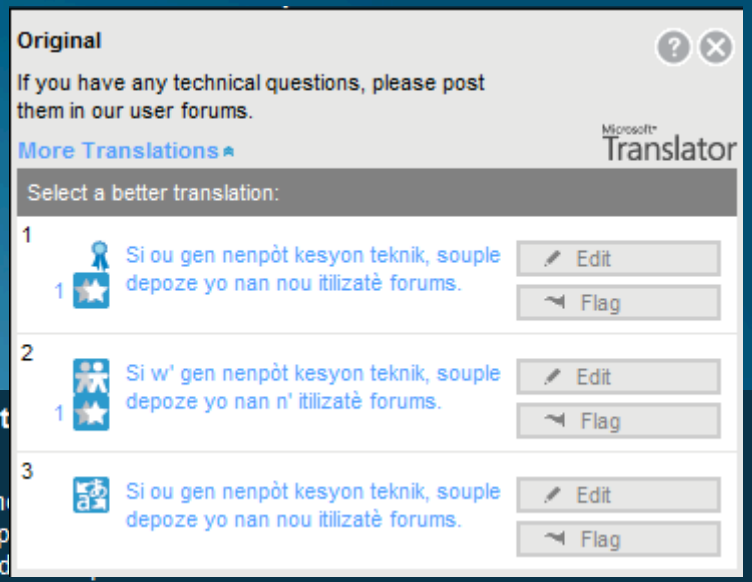
Tools Available for Haitian Creole

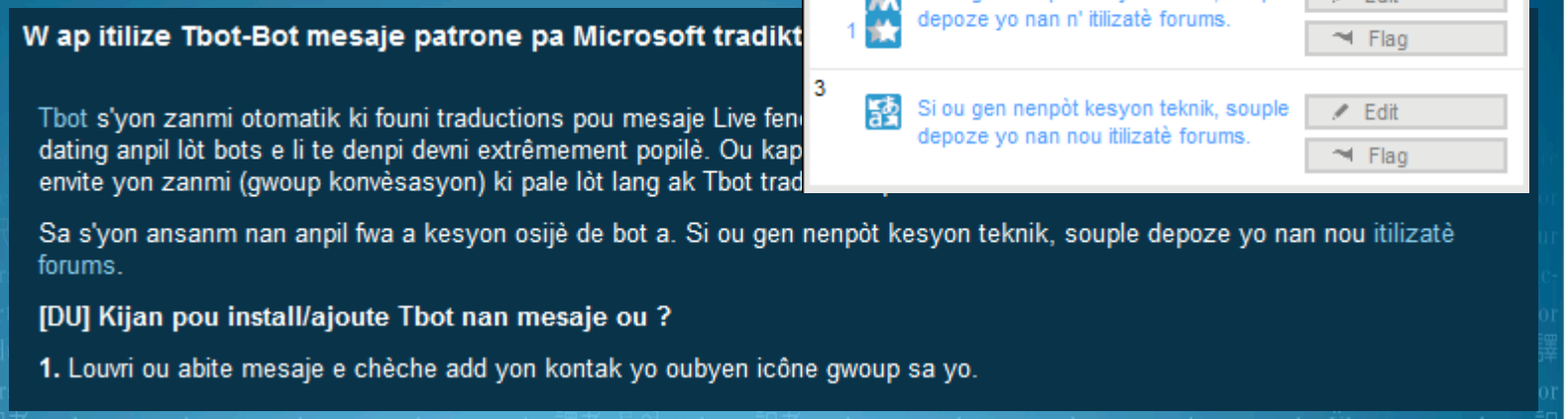
- Home page (Web page viewer, cut-and-paste translator)
- Haitian Creole one of the languages available through our API (Advanced Programming Interface)
  - Multiple interfaces: AJAX, SOAP, HTTP
  - Can integrate translation directly into a variety of apps
- Widget
  - Integrate translation into Web pages
  - Traffic kept client side



Tools Available for Haitian Creole

- Widget/Collaborative Translation Framework (CTF)
  - Community can contribute translations
  - These can be published to Web pages
  - Mixes MT with “trusted” human translations

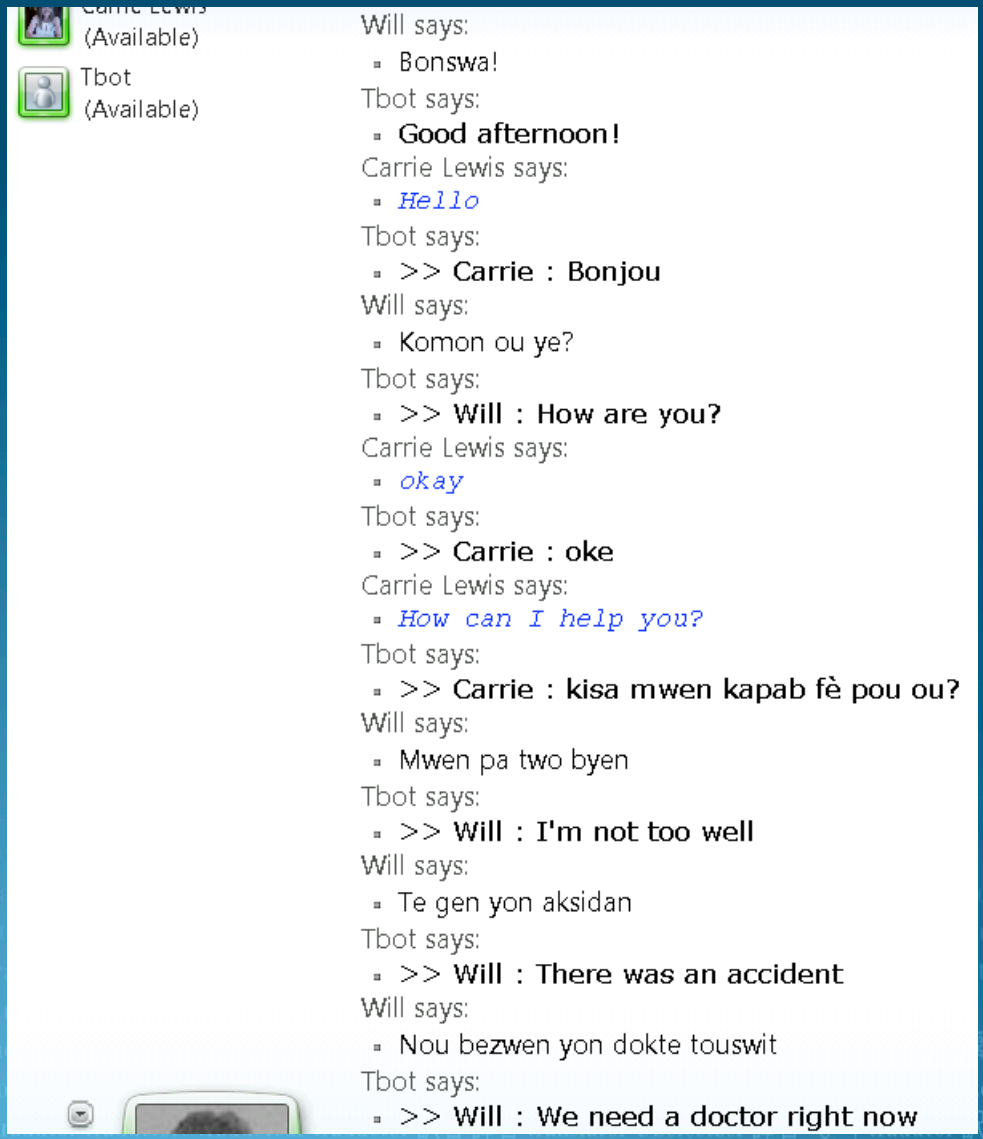




Demo of Widget/CTF

Real Time IM Translation

- T-Bot: Provides real-time translations of IM
  - Add as a participant
  - Translates between the languages selected
- SMS content in training probably helps with IM



Overview

- Earthquake in Haiti created a significant humanitarian crisis
- NLP/MT technology can be useful in such crises
- MT can be developed for low-data languages
- Such MT can be rolled out quickly, even in a production environment, and even when starting with very little
- Critical problem for any Low-Resource Language: Data
  - In Haitian crisis, barriers to data access were lowered
  - Many participants donated data in addition to time
  - Preemptive work for other low-data languages may require data sharing agreements
  - Large-scale data sharing *a la* TAUS TDA may help in low data language tool and resource development

Public API V2 Sample Code

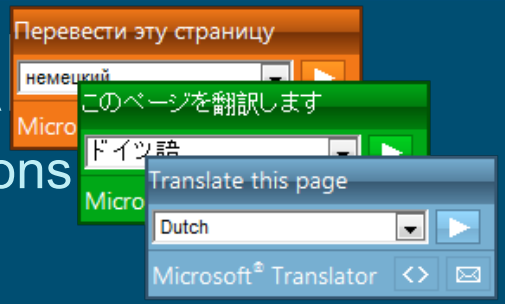
```
TranslateOptions options = new TranslateOptions();
        options.SentenceLengths = true;
        options.Uri = "www.foo.com";
        options.MaxTranslations = 4;
        options.Category = "general";
        options.ContentType = "text/plain";
        options.User = "Rachel";

        string[] texts = new string[2];
        texts[0] = "this is my first one";
        texts[1] = "this is my second one";

        TranslateResponse response =
_soapClient.TranslateArray(_appId, texts, "en", “ht", options);
```

Translator Widget & AJAX A

Enables any website to provide instant, in-place translations



- Simple copy/paste of widget code snippet
- Gives webmasters control of their translation UX