CS11-737 Multilingual NLP

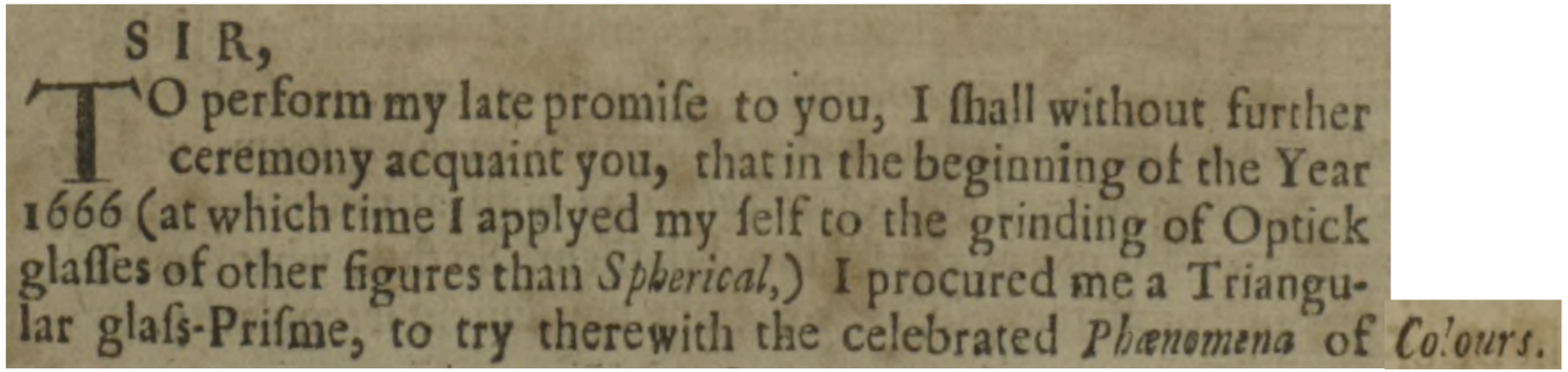
Language Change, Pidgin, Creoles

Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

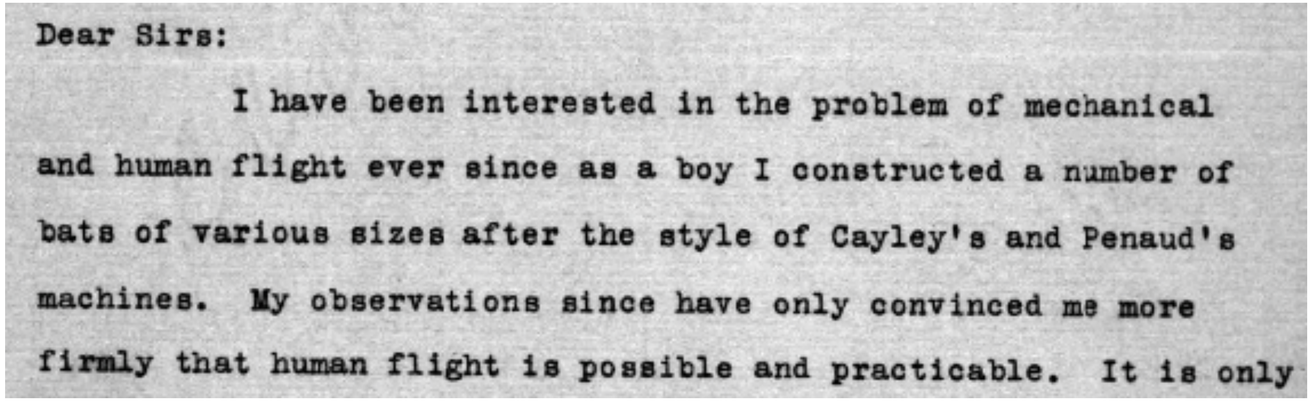
Language is changing!



SIR,

TO perform my late promiſe to you, I ſhall without further ceremony acquaint you, that in the beginning of the Year 1666 (at which time I applyed my ſelf to the grinding of Optick glaſſes of other figures than *Spherical*,) I procured me a Triangular glaſs-Priſme, to try therewith the celebrated *Phænomena* of *Colours*.

Letter from Isaac Newton in 1672.



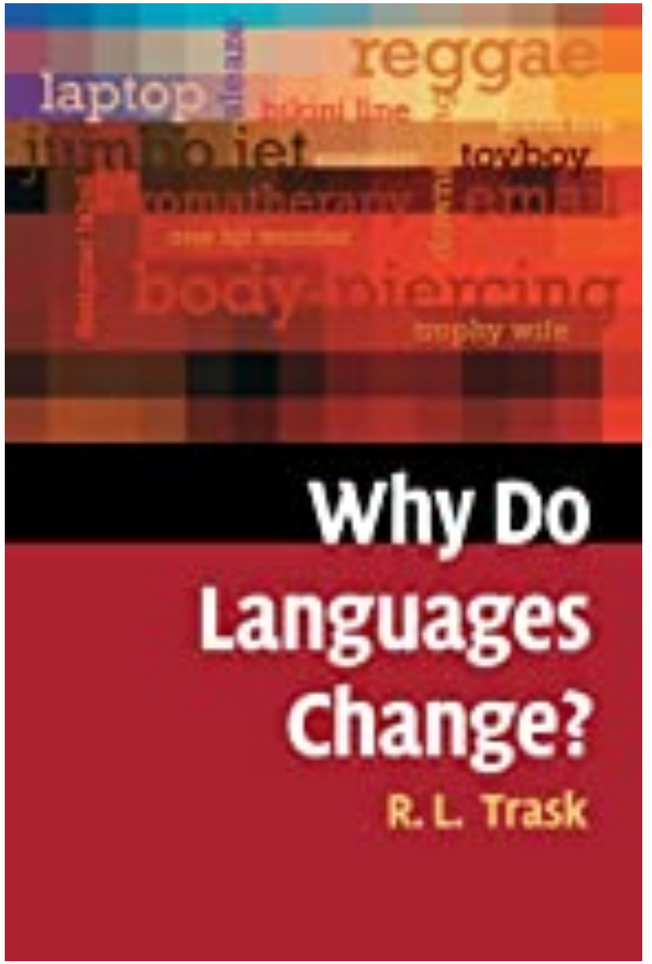
Dear Sirs:

I have been interested in the problem of mechanical and human flight ever since as a boy I constructed a number of bats of various sizes after the style of Cayley's and Penaud's machines. My observations since have only convinced me more firmly that human flight is possible and practicable. It is only

Letter from Wilbur Wright 1899

Why do languages change?

- Changes in the world
  - ∅ -> email, radiogram -> ∅
- Laziness/efficiency (Gibson 2019)
  - telephone -> phone
- Emphasis/clarity
  - he/heo/hi -> he/she/they
- Politeness
  - https://developers.google.com/style/word-list
- Misunderstanding
  - bead: prayer -> small ball
- Group identity/prestige (Danescu-Niculescu-Mizil et al. 2013)
  - aroma -> smell
- Structural reasons
  - regularity in phonetics, morphology



(Trask 2010)

Cognates

https://www.mentalfloss.com/article/68281/evolution-two-indo-european-language-family

Loan Words

orchestra

オーケストラ

カラオケ

"empty - orche"

karaoke

Loan Words in Chinese

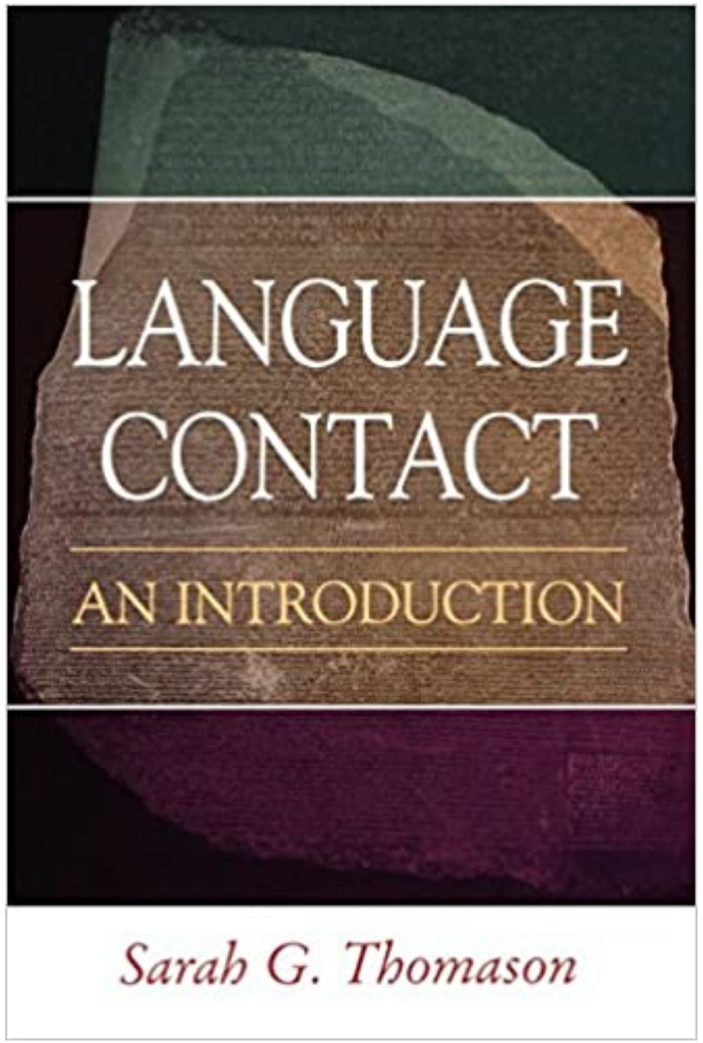
Sanskrit	pron.		Chinese	pron.	meaning
क्षण	kṣaṇa	→	剎那	chànà	instant
बिम्बा	bimbā	→	苹果	píngguǒ	apple

English		Chinese	pron.
coffee	→	咖啡	kāfēi
t-shirt	→	T恤衫	

Language contact

- Language contact is the use of more than one language in the same place at the same time (Thomason ‘95)

- Major driving factor behind language change



Arabic--Swahili

- Swahili - major language in southeast Africa, 100M speakers
- 800 A.D.-1920 Indian Ocean trading
- Influence of Islam
- ~40% of Swahili types are borrowed from Arabic (Johnson ‘39)

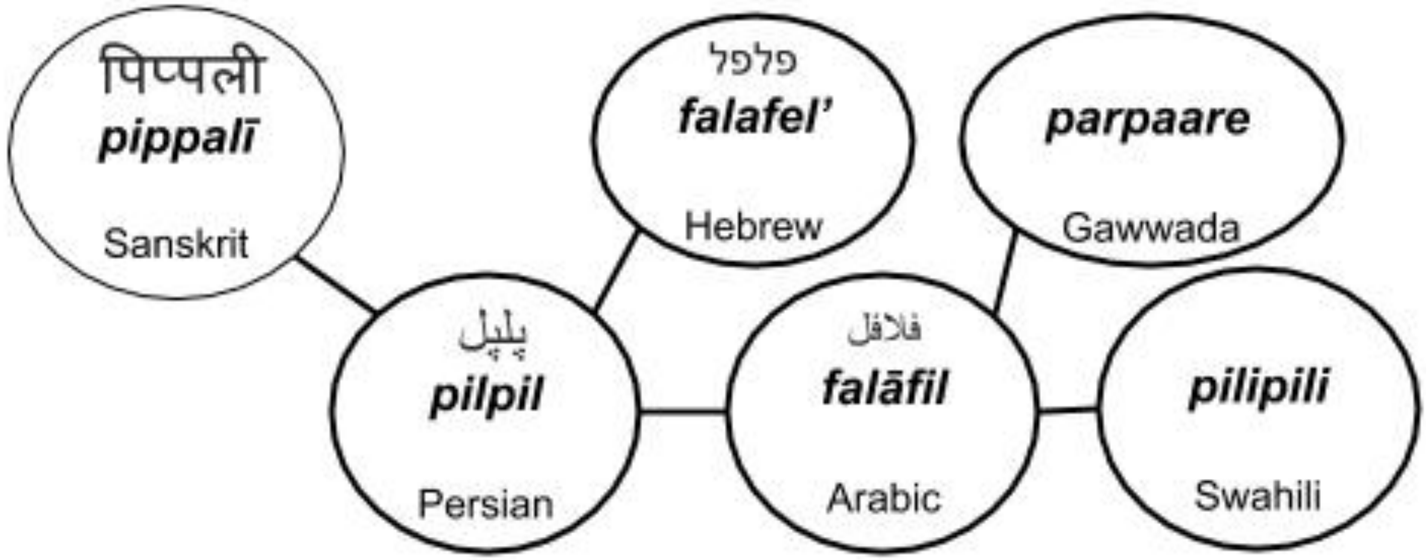
Lexical borrowing is pervasive in languages

Resource-poor recipient	**# speakers (millions)**	**Resource-rich donors (% types)**
Swahili, Zulu, Malagasy, Hausa, Tarifit, Yoruba	200	Arabic, Spanish, English, French (>40%)
Japanese, Vietnamese, Korean, Cantonese, Thai	400	Chinese, English (30–70%)
Hindustani, Hindi, Urdu, Bengali, Persian, Pashto	860	Arabic, English (>40%)
	1.4 billion	

Cross-lingual lexical similarities

- How to bridge across languages?

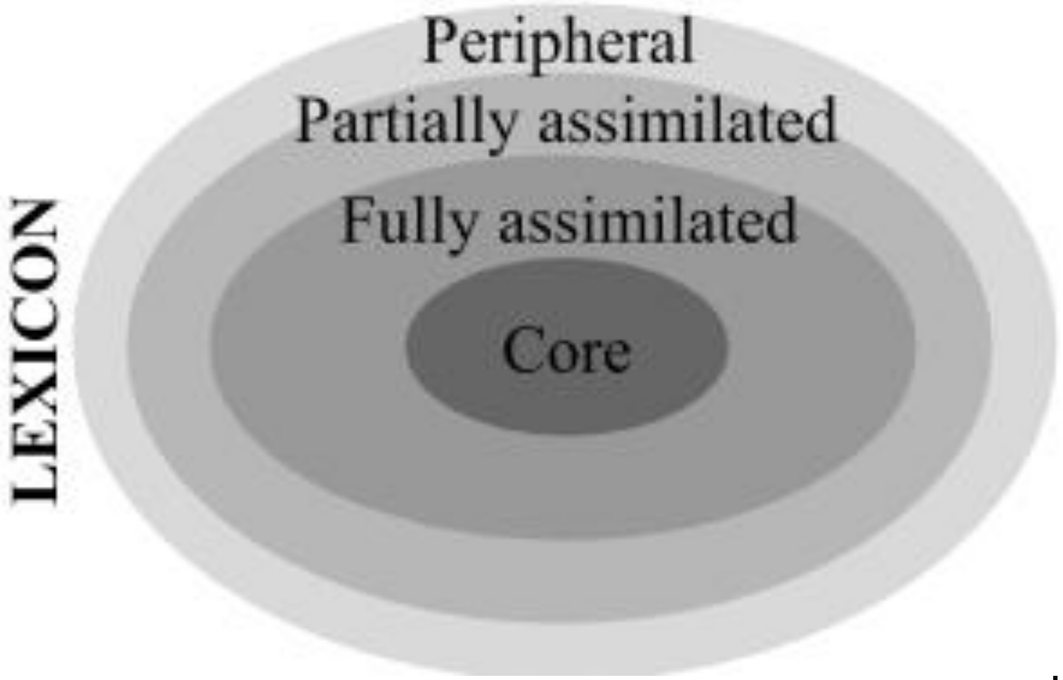
- Identify words that are orthographically or phonetically similar across different languages and are likely to be mutual translations



Lexicon structure

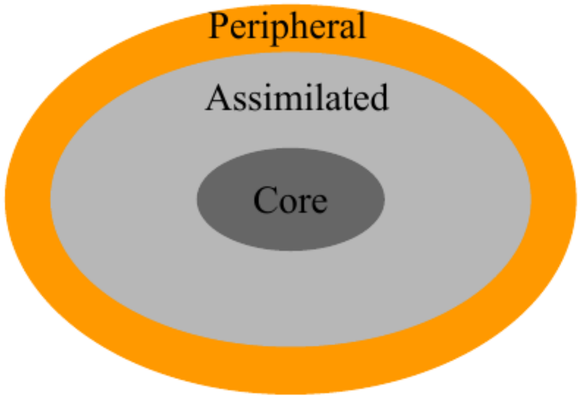
- Core-periphery lexicon structure (Itô & Mester ‘95)

- English:
  - Core (20%–33%): *beer, bread*
  - Assimilated: *cookie, sugar, coffee, orange*
  - Peripheral: *New York, Luxembourg*



How to bridge across languages?

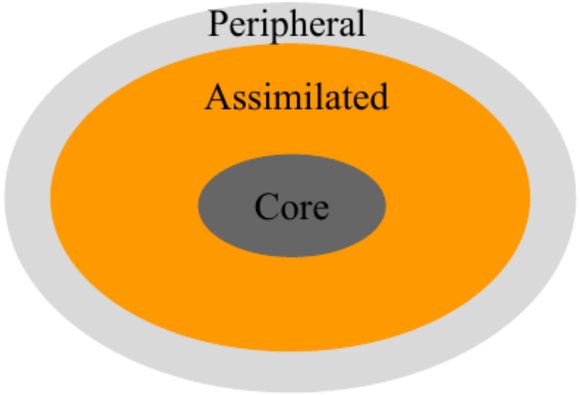
Transliteration



Peripheral vocabulary:
proper names, specialized terms

English	New York
Yoruba	Niu Yoki
Russian	Нью-Йорк
Arabic	نيويورك
Hebrew	ניו יורק

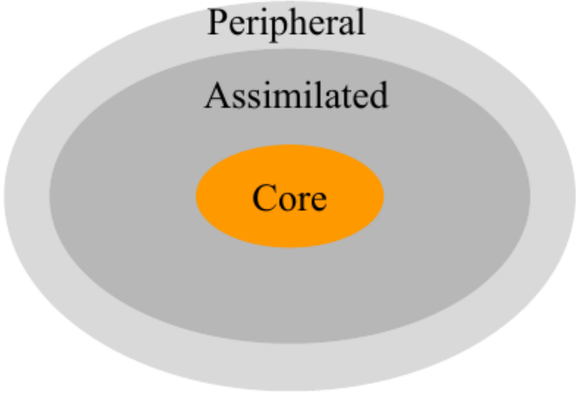
Borrowing



Content words of foreign origin, assimilated in the language and aren't perceived as foreign

Arabic	سكر
*transliterated	sukkar
Latin	zuccarum
French	sucre
German	Zucker
Italian	zucchero
English	sugar

Cognates

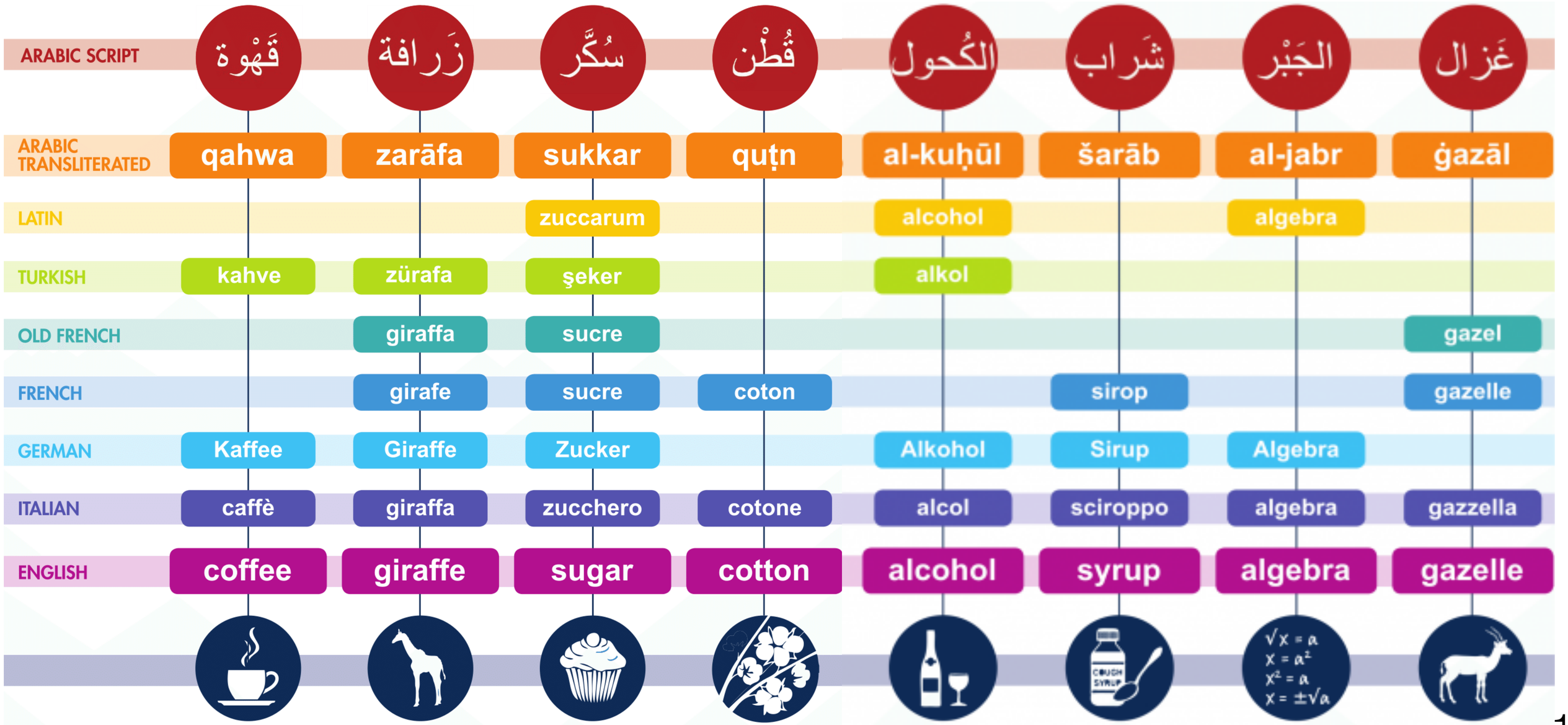


Content words in core lexicon: words in related languages inherited from one word in a common ancestral language

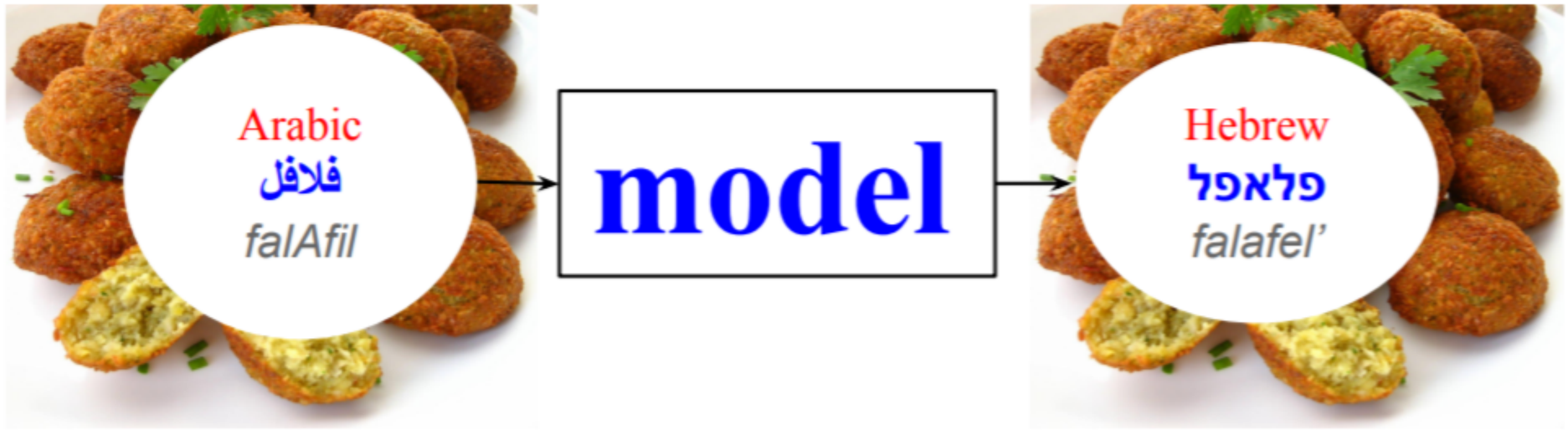
Latin	nocte
French	nuit
Spanish	noche
Italian	notte
Portugese	noite
Romanian	noapte

Cross-lingual Lexical Learning

Mapping lexicons across languages

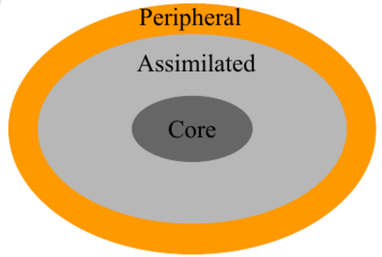


Cross-lingual lexicon induction



Transliteration models

- FSTs Knight & Graehl ‘98
- LSTMs with attention Rosca & Breuel’16
- Exact Hard Monotonic Attention for Character-Level Transduction Wu & Cotterell’19



Peripheral vocabulary:
proper names, specialized terms

English	New York
Yoruba	Niu Yoki
Russian	Нью-Йорк
Arabic	نيويورك
Hebrew	ניו יורק

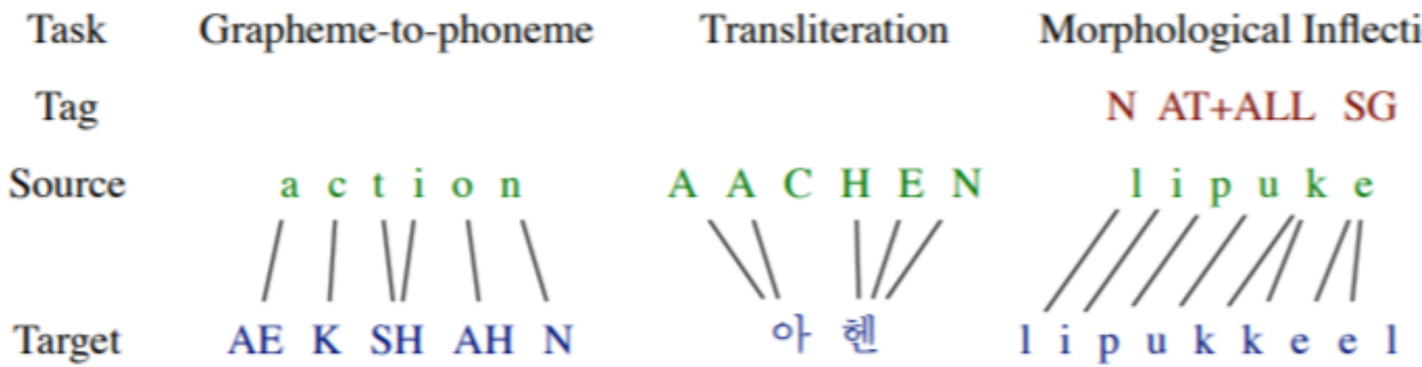


Figure 1: Example of source and target string for each task. Tag guides transduction in morphological infl

Transliteration evaluation

Intrinsic evaluation

- Word accuracy in top-1
- Fuzziness in top-1 (mean F-score)
- Ranking; Mean Reciprocal Rank (MRR), Mean Average Precision (MAP)

Report of NEWS 2018 Named Entity Transliteration Shared Task

Nancy Chen[1], Rafael E. Banchs[2], Min Zhang[3], Xiangyu Duan[3], Haizhou Li[4]

Downstream evaluation

- Machine translation
- Cross-lingual information extraction

Transliteration resources

- 1.6M named entities across 180 languages aggregated across multiple public datasets

TRANSLIT: A Large-scale Name Transliteration Resource

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, Mark Cieliebak
Zurich University of Applied Sciences, Deep Impact
Switzerland
benf@zhaw.ch, gilbert@deep-impact.ch, vode@zhaw.ch, ciel@zhaw.ch

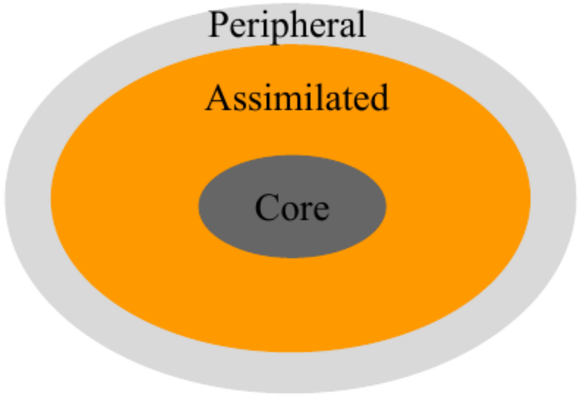
Abstract

Transliteration is the process of expressing a proper name from a source language in the characters of a target language (e.g. from Cyrillic to Latin characters). We present TRANSLIT, a large-scale corpus with approx. 1.6 million entries in more than 180 languages with about 3 million variations of person and geolocation names. The corpus is based on various public data sources, which have been transformed into a unified format to simplify their usage, plus a newly compiled dataset from Wikipedia.
In addition, we apply several machine learning methods to establish baselines for automatically detecting transliterated names in various languages. Our best systems achieve an accuracy of 92% on identification of transliterated pairs.

Keywords: Transliteration of Names, Name Variant Discovery, Multi-lingual, Language Resource

Cognates and loanwords

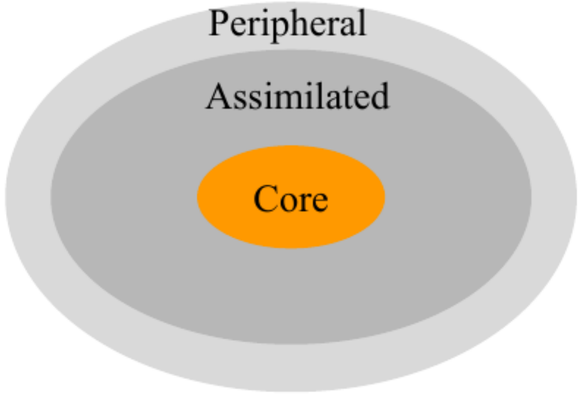
Borrowing



Content words of foreign origin, assimilated in the language and aren't perceived as foreign

Arabic	سكر
*transliterated	sukkar
Latin	zuccarum
French	sucre
German	Zucker
Italian	zucchero
English	sugar

Cognates



Content words in core lexicon: words in related languages inherited from one word in a common ancestral language

Latin	nocte
French	nuit
Spanish	noche
Italian	notte
Portugese	noite
Romanian	noapte

Arabic--Swahili borrowing examples

English	**Arabic** Semitic	**Swahili** Bantu	**Phonological & morphological integration**
fever	حمى ḥummat	homa	* syllable structure adaptation: CV, CVV, CVC, CVCC → V, CV * degemination - Swahili does not allow consonant clusters * vowel substitution
minister	الوزير Alwzyr	kiuwaziri	* Arabic morphology (optionally) drops * Swahili morphology is applied * vowel epenthesis to keep syllables open * vowel substitution
palace	القصر AlqSr	kasiri	* consonant adaptation: /tˤ/→/t/, /dˤ/→/d/, /θ/→/s/, /x/→/k/, etc * vowel epenthesis

Linguistic research on lexical borrowing

- Case studies of lexical borrowing in language pairs
  - Cantonese (Yip ‘93), Korean (Kang ‘03), Thai (Kenstowicz & Suchato ‘06), Russian (Benson ‘59), Romanian (Friesner ‘09), Hebrew (Schwarzwald ‘98), Yoruba (Ojo ‘77), Swahili (Schadeberg ‘09), Finnish (Johnson ‘14), 40 languages (Haspelmath & Tadmor ‘09), etc.

- Case studies of phonological/morphological phenomena in borrowing
  - Phonological integration (Holden ‘76, Van Coetsem ‘88, Ahn & Iverson ‘04, Kawahara ‘08, Hock & Joseph ‘09, Calabrese & Wetzels ‘09, Kang ‘11); morphological integration (Rabeno ‘97, Repetti ‘06); syntactic integration (Whitney ‘81, Moravcsik ‘78, Myers-Scotton ‘02), etc.

- Case studies of sociolinguistic phenomena in borrowing
  - (Guy ‘90, McMahon ‘94, Sankoff ‘02, Appel & Muysken ‘05), etc.

Cognate and loanword models

- Phonologically-weighted Levenshtein distance between phonetic sequences Mann & Yarowsky ‘01, Dellert ‘18
- Phonetic + semantic distance Kondrak ‘01, Kondrak,Marcu & Knight ‘03
- Log-linear model with Optimality-theoretic features Bouchard-Côté et al. ‘09
- Generative models of sound laws and word evolution for cognate identification Hall & Klein ‘10, ‘11
- Optimality-theoretic constraint-based learning for loanword identification Tsvetkov & Dyer ‘16
- Cognate identification using Siamese networks Soisalon-Soininen & Granroth-Wilding ’19

Cognate databases

- 3.1 million cognate pairs across 338 languages using 35 writing systems

CogNet: a Large-Scale Cognate Database

Khuyagbaatar Batsuren[†] **Gábor Bella**[†] **Fausto Giunchiglia**[†§]
DISI, University of Trento, Trento, Italy[†]
Jilin University, Changchun, China[§]
{k.batsuren; gabor.bella; fausto.giunchiglia}@unitn.it

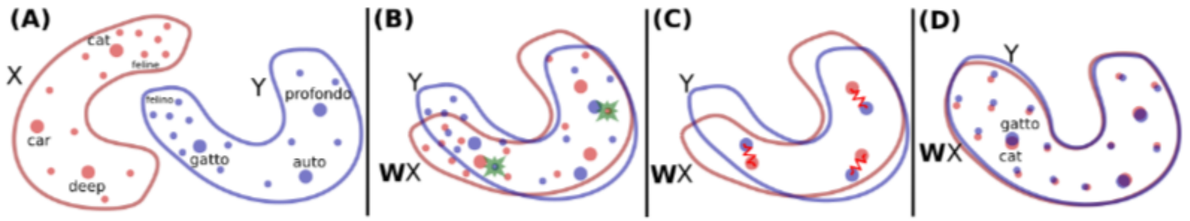
Lexical borrowing databases

https://wold.clld.org/

Bilingual lexicon induction

1. Learn monolingual embeddings
2. Find alignment between embedding spaces
3. Find nearest neighbors to induce lexicon
4. Perform supervised alignment to minimize distance between lexicon items

MUSE: Multilingual Unsupervised and Supervised Embeddings



https://ruder.io/cross-lingual-embeddings/

Code-switching

- Usually between speakers of equal fluency in two languages
  - Casual speech/text
  - Face-to-face or in social media
- Often defined in terms of a matrix language (based language word order)
  - (major) Word order in matrix languages, as are particles (morphology)
  - Lexemes and short phrases in other language
- We can measure the amount of code switching
  - As percentage (but doesn't distinguish number of switchers)
  - Other measure try to capture that (but its hard)
-

Pidgins and Creoles

- Pidgins: non-natives learn a “simplified” mix of multiple languages
- Creoles: when such a mix/non-native dialect becomes a native languages
  - Creoles have native speakers, Pidgins do not (yet-ish).
  - Might be classified as just dialects, possibly low-status
- Jamaican Patois (Creole)
  - An fram Dievid taim op tu wen dem did tek we di Izrel piipl dem an fuos dem fi go wok a Babilan a fuotiin jinarieshan Jiizas did av de-so tu, an fram da taim de tu wen Krais Jiizas bi pan, a fuotiin jinarieshan dat tu.
- Haitian Creole
  - Zebadya pitit Izmayèl la chèf branch fanmi jida a va pi gwo jij pou tou sa ki an rapò ak lalwa peyi a.
- English
  - Is a creole or code-switched Saxon and Norman French (arguably)

Pidgins and Creoles

- These are linguistic terms
- They also get used as political terms too
  - And not always favorably
  - “not proper English”, “dialect”, “uneducated speech”
- Often used for speech communication not in writing
  - Cf Latin verses Vulgar Latin
- Thus often hard to find examples, written forms are formal
- But what about code-switching is it a “pre-pidgin”?

Major Code-switching Dialects

- Sometimes between local and global languages
  - But most common examples are between major languages
- Hinglish
  - Hindi and English: very common with educated young Hindi speakers
- Chinglish
  - Chinese and English: common amongst Chinese speakers in Singapore
- Spanglish
  - Spanish and English: common in Spanish speaking areas of US (e.g. South West, New York, Puerto Rico)
- African American English and Standard American English
  - Very common in US in Black communities
- In these cases people are usually very fluent in both languages

When do they code-switch

- Vocabulary coverage:
  - Talk about machine learning in English and food in Hindi
  - Relationships in Spanish, studying in English
- “Showing off” Trends
  - “Fashionable English Works” vs “Ethnic other words”
  - Displaying affiliation (show education and/or show roots)
- Maybe more sentiment in native languages (Rudra et al 2016)
  - Looking at language choice in tweets
- Entrainment (copy others in conversation)
- To get distinctions (simple semantics or opinion)

Why do we care?

- Newpapers and Wikipedia will not be code-switched
  - Why care about NLP of non-standard language?
- Code-switching is how people actually communicate
- Code-switching
  - People type in questions to Google/Bing
  - They talk to call centers
  - They write their opinions
  - Use of code-switching can define group membership
  - People trust code-switched communication better
    - (Not that it is fake, but its their language and someone developed communication in their language)
- Facebook, Amazon, Microsoft, Apple all want to understand Code-switch more

Why is it Hard

- There is very little data available
  - Often code-switched data is removed from datasets
- It is very noisy
  - Spelling is very non-standard so its hard to know the vocabulary
  - It often romanizes native script (inconsistently)
- Our favorite contextualized word embeddings are confused
  - We have “random” mixed language juxtaposed
  - BERT was never trained for that
  - mBERT was never trained for that
- Its casual speech/text
  - Monolingual casual speech is hard, we now have two languages

Code Switching Data

- Often used by hard to find
  - Harder to verify: is it bilingual or code switched
- Twitter/youtube/reddit
  - Social media is good, but its not labeled (and very noisy)
- Collecting is hard too
  - Need right environment to have users code-switch
  - Usually based on their peer relationship.
  - There isn’t just one type of code-switching

Code Switching Data

- Annual speech/text workshops on Code-Switching
  - At ACL/Interspeech
- Sitaram et al. “A survey of code-switched speech and language processing”
  - https://arxiv.org/abs/1904.00784
- Thamar Solorio (U Huston)
  - https://ritual.uh.edu/code-switching/
- Growing but still limited
  - Only a few language pairs are studies

Code Switching: LT Tasks

- Language ID (Speech ID)
  - Labeling the words
  - In Speech the pronunciation may vary from monolingual cases)
- Speech Recognition/Synthesis
  - Google's Indian English ASR is actually Hinglish ASR
  - Code-switch synthesis, but ...
- Spelling Normalization
  - As spell checkers don't work, spelling is particularly inconsistent
  - May require roman → native script conversion too
- POS Tagging
  - Some datasets available (but often noisy)
  - Detecting matrix language can be important

Code Switching: LT Tasks

- Named Entity Recognition
  - (and cross-lingual entity linking)
  - Various challenges have addressed this
- Sentiment analysis
  - Understanding cross lingual references may be important
- Question Answering
- NLI
- Dialog processing
  - Common Amigos (Ahn et al 2020, Parekh et al 2020)
- When/home to use code-switching
  - Generation as a style

Code Switching: Techniques

- Very similar to general low-resource language issues
  - Find appropriate data
  - Bootstrap labeling
  - Data argumentation/generation techniques
  - Find new (reliable) evaluation techniques

Mining Data

- Social media sites
  - Code Switching (usually) implies casual speech
- Youtube for speech
  - Hinglish: lots of examples, but some human has to find each video
    - Google Indian English ASR can give a reasonable transcription
  - Broadcast news, Bibles/Koran wont have code-switching
- Reddit (or local equivalent)
  - Mining the data is hard
  - You want conversations, not just utterances
- Twitter/Weibo
  - Rarely conversational
  - Sometimes bilingual (translations) not code switching

Bootstrapping Labeling

- Label small amount of data
  - Build classifier for the data
  - Use the classifier to label lots of other data
  - Select “high confidence” samples and add to training data
  - Rebuild classifier
  - Repeat until (something)
- Care has to made to ensure you don’t miss out on important examples
- Care has to made to ensure you don’t just add garbage examples
- Care has to made to ensure you don’t just add trivial addition examples
- Downstream task evaluation would help
  - But you probably don’t have that yet.

Akshat Gupta, Sargam Menghani, Sai Krishna Rallabandi and Alan W Black "Unsupervised Self-Training for Sentiment Analysis of Code-Switched Data", Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching 2021

Data augmentation/generation

- Build generator from limited data to get more
  - Build classifiers that distinguish real from generated data
  - Choose false positive data to boost base data
  - Been shown to help in lots of cases
  - (could this help building better word embedding models)

- Paraphrase existing data to get more
  - Replace NE, modify some word by translation

Evaluation Techniques

- Task evaluation of held out data
  - Standard techniques, but does that help the overall task
- We don’t yet have lots of good high level tasks to test
  - Dialog understanding
  - Summarization
  - Question/Answering
- MSR India (Khanuja et al ACL 2020)
  - GLUECoS: Set of standard tasks for testing code-switching models
  - https://arxiv.org/abs/2004.12376

Code-switching, Pidgins and Creoles

- Multilingual is much more varied than one single language
- Code-switching
  - Mixed within an utterance
- Pidgin
  - Non-native mixed lingua franca (often for trade)
- Creole
  - When Pidgin becomes native and its own language
- Issues are as hard as with monolingual casual speech
  - But now we have multiple languages to confuse things

Discussion

Class discussion

- Pick a language that you speak, read about its history, and in particular how this language influenced other languages
  - are there languages that historically borrowed words from your language?
  - can you find specific examples of words?
  - could you recognize these loanwords in other languages based on their new form?
  - can you guess what were phonological and morphological adaptation processes that the loanword had to undergo to assimilate in the new language?
  - Pick another language and analyze the code-switching