

CS11-737: Multilingual Natural Language Processing

Language contact and change

Graham Neubig

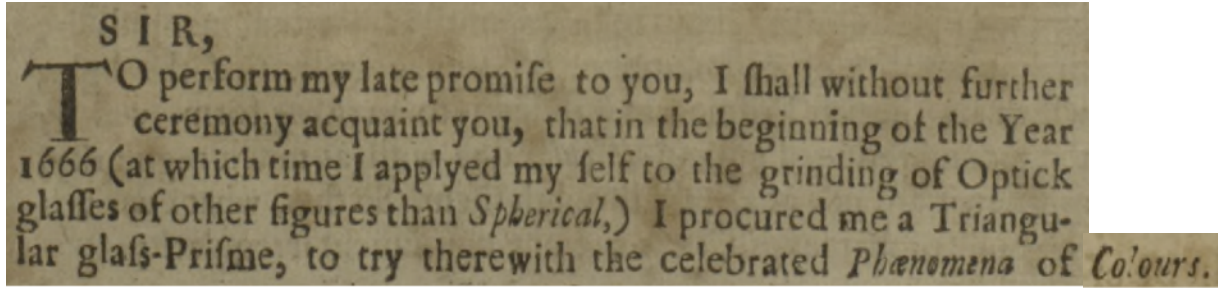
(Many Slides by Yulia Tsvetkov)



Carnegie Mellon University

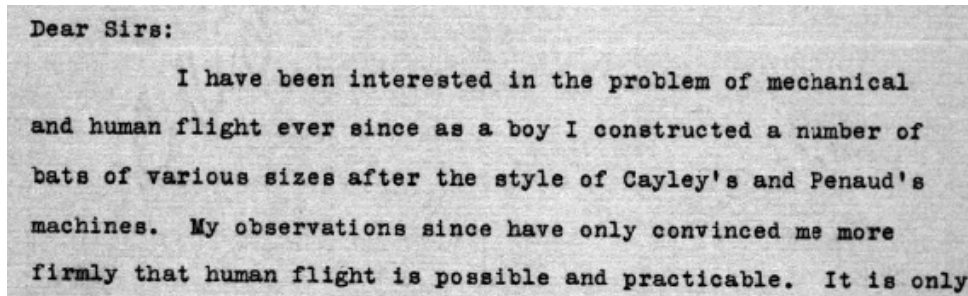
Language Technologies Institute

Language is changing!



S I R,
T O 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 glassees of other figures than *Spherical*;) I procured me a Triangular glasse-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

Hello,

[This is an automated response]

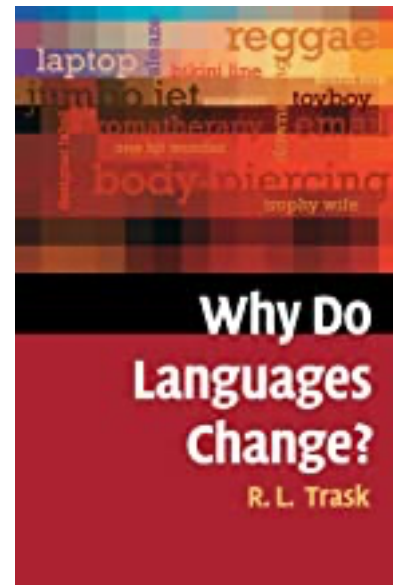
Thank you for your email. I'll be away from the office from December 18 - January 2 for winter holidays! My email responses may be sparse during this time. I will try to follow up after I'm back at work, but if you don't get a response please feel free to follow up sometime after I get back.

Graham

Email from Graham Neubig 2021

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)

Lexical Changes: Cognates and Loanwords

Cognates



Loan Words

orchestra



オーケストラ



karaoke



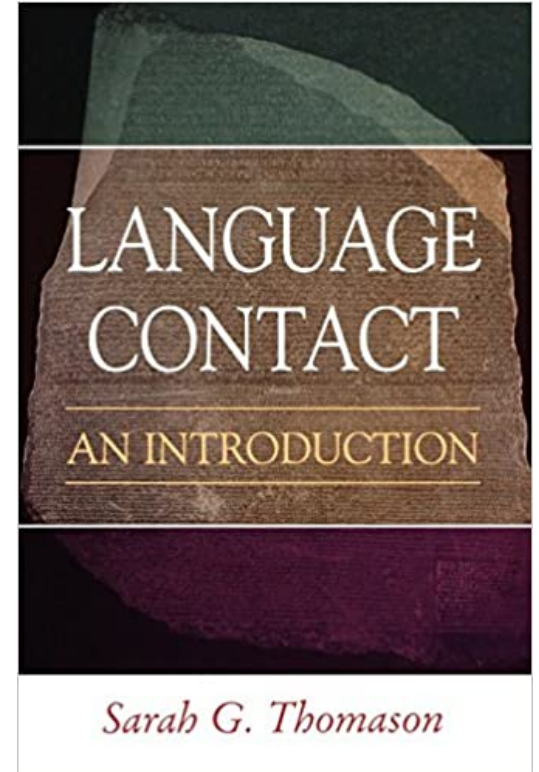
↓ カラオケ
"empty - orche"



Language Contact and the Lexicon

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)



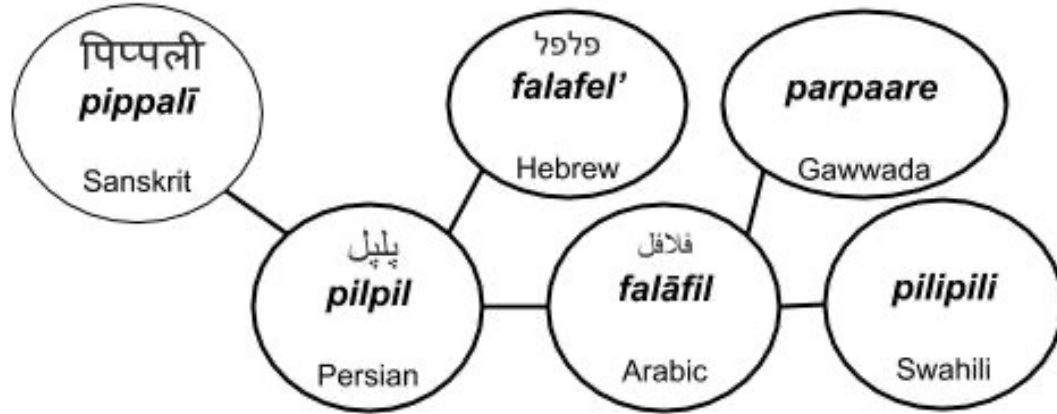
Lexical borrowing is pervasive in languages

| 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 | |

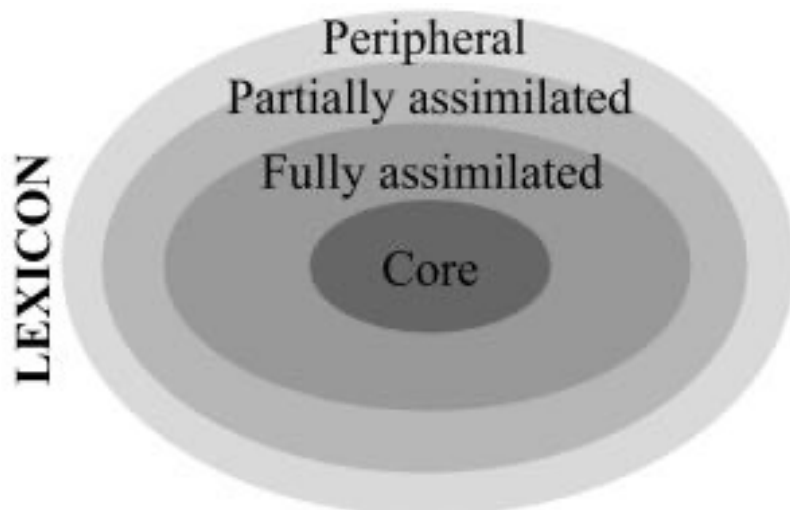
- Not by chance! Resources associated with reach/social influence

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



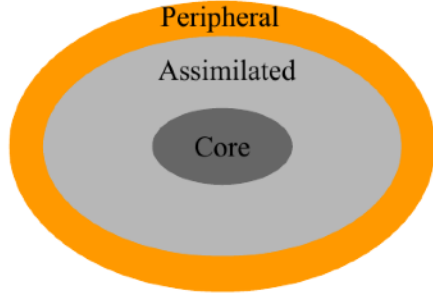
Lexicon structure



- 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?

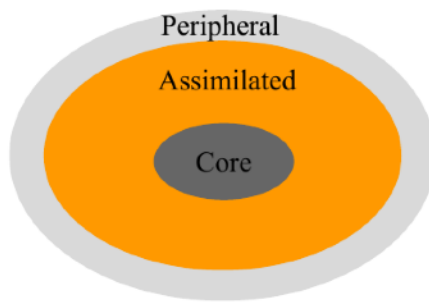
Transliteration



Peripheral vocabulary:
proper names, specialized terms

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

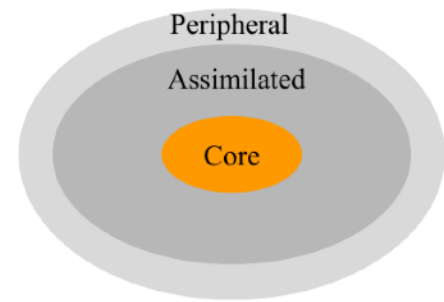
Borrowing



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

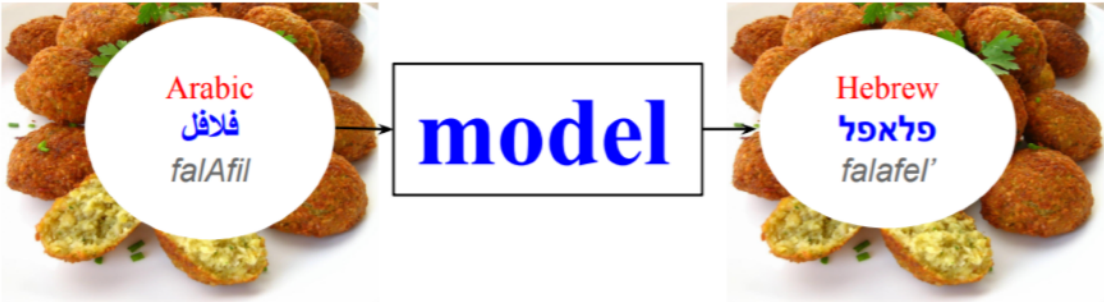


Content words in core lexicon:
words in related languages
inherited from one word in a
common ancestral language

| | |
|------------|--------|
| Latin | nocte |
| French | nuit |
| Spanish | noche |
| Italian | notte |
| Portuguese | noite |
| Romanian | noapte |

Cross-lingual Lexical Learning

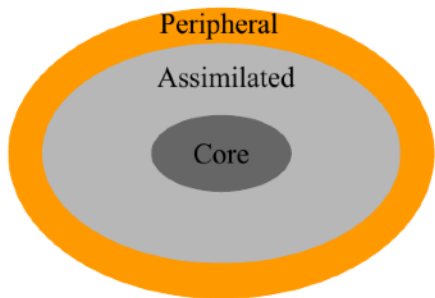
Cross-lingual lexicon induction



>100K words

A vertical list of words, likely representing a large vocabulary. The words are arranged in two columns, with a dashed arrow pointing upwards on the right side. One word in the list is circled in red, indicating its significance in the context of the study.

Transliteration models



Peripheral vocabulary:
proper names, specialized terms

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

- FSTs [Knight & Graehl '98](#)
- LSTMs with attention [Rosca & Breuel'16](#)
- Exact Hard Monotonic Attention for Character-Level Transduction [Wu & Cotterell'19](#)

| Task | Grapheme-to-phoneme | Transliteration | Morphological Inflection |
|--------|---------------------|-----------------|--------------------------|
| Tag | | | N AT+ALL SG |
| Source | a c t i o n | A A C H E N | l i p u k e |
| Target | AE K SH AH N | 아 헨 | l i p u k k e e l l e |

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

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 Chen¹, Rafael E. Banchs², Min Zhang³, Xiangyu Duan³, Haizhou Li⁴

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

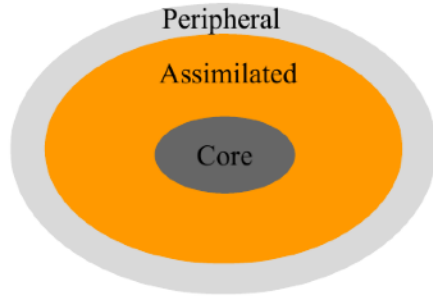
Transliteration is the process of expressing a proper name from a source language in the characters of a target language (e.g. from Cyrillic to Latin characters). We present TRANSLIT, a large-scale corpus with approx. 1.6 million entries in more than 180 languages with about 3 million variations of person 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

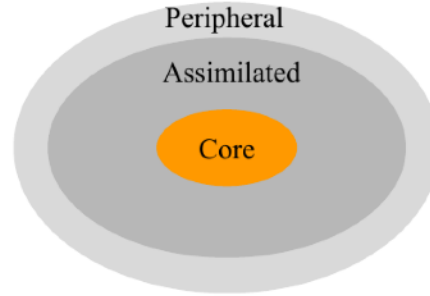
Borrowing



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 | <ul style="list-style-type: none">* syllable structure adaptation: CV, CVV, CVC, CVCC → V, CV* degemination - Swahili does not allow consonant clusters* vowel substitution |
| minister | الوزير Alwzyr | kiuwaziri | <ul style="list-style-type: none">* Arabic morphology (optionally) drops* Swahili morphology is applied* vowel epenthesis to keep syllables open* vowel substitution |
| palace | القصر AlqSr | kasiri | <ul style="list-style-type: none">* 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

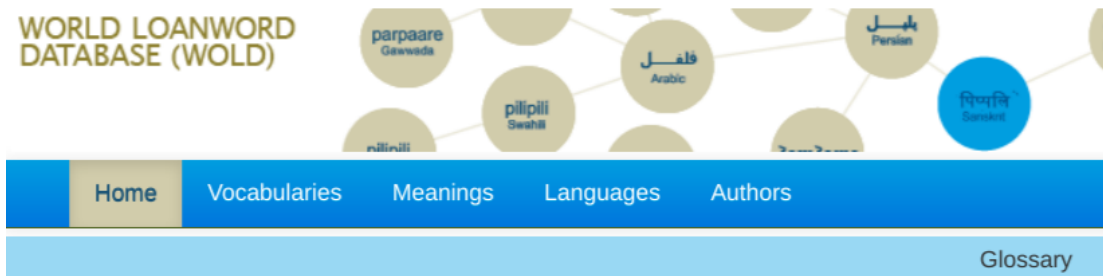
Khuyagbaatar Batsuren[†] Gábor Bella[†] Fausto Giunchiglia^{†§}

DISI, University of Trento, Trento, Italy[†]

Jilin University, Changchun, China[§]

`{k.batsuren; gabor.bella; fausto.giunchiglia}@unitn.it`

Lexical borrowing databases



The World Loanword Database (WOLD)

The World Loanword Database, edited by [Martin Haspelmath](#) and [Uri Tadmor](#), is a scientific publication by the [Max Planck Institute for Evolutionary Anthropology](#), Leipzig (2009).

It provides [vocabularies](#) (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word. It allows users to find [loanwords](#), [source words](#) and [donor languages](#) in each of the 41 languages, but also makes it easy to compare loanwords across languages.

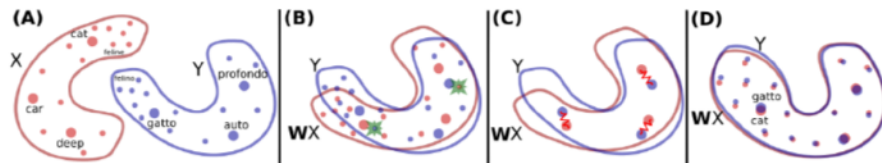
Each vocabulary was contributed by an expert on the language and its history. An accompanying book has been published by De Gruyter Mouton ([Loanwords in the World's Languages: A Comparative Handbook](#), edited by [Martin Haspelmath & Uri Tadmor](#)).

<https://wold.clld.org/>

Bilingual lexicon induction

- (A) Learn monolingual embeddings
- (B) Find alignment between embedding spaces
- (C) Find nearest neighbors to induce lexicon
- (D) Perform supervised alignment to minimize distance between lexicon items

MUSE: Multilingual Unsupervised and Supervised Embeddings



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

Discussion

Class discussion

- **Option 1:** Read "[How Efficiency Shapes Human Language](#)". Think about a language you speak. What are some elements of this language that you think are efficient, and some that you think are inefficient?
- **Option 2:** 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?