Language contact

Yulia Tsvetkov
Language contact

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

Factors driving the change of languages and language varieties:

- **Language-internal**
  - ease of articulation
  - analogy/reinterpretation
  - language contact

- **Language-external**
  - language contact
  - geography
  - social prestige
    - conscious
    - subconscious
Arabic--Swahili

- 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

<table>
<thead>
<tr>
<th>Resource-poor recipient</th>
<th># speakers (millions)</th>
<th>Resource-rich donors (% types)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swahili, Zulu, Malagasy, Hausa, Tarifit, Yoruba</td>
<td>200</td>
<td>Arabic, Spanish, English, French (&gt;40%)</td>
</tr>
<tr>
<td>Japanese, Vietnamese, Korean, Cantonese, Thai</td>
<td>400</td>
<td>Chinese, English (30–70%)</td>
</tr>
<tr>
<td>Hindustani, Hindi, Urdu, Bengali, Persian, Pashto</td>
<td>860</td>
<td>Arabic, English (&gt;40%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.4 billion</td>
</tr>
</tbody>
</table>
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.
<table>
<thead>
<tr>
<th>Arabic Script</th>
<th>ARABIC TRANSLITERATED</th>
<th>LATIN</th>
<th>TURKISH</th>
<th>OLD FRENCH</th>
<th>FRENCH</th>
<th>GERMAN</th>
<th>ITALIAN</th>
<th>ENGLISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>قهوة</td>
<td>qahwa</td>
<td>zuccarum</td>
<td>kahve</td>
<td>giraffa</td>
<td>girafe</td>
<td>Kaffee</td>
<td>caffè</td>
<td>coffee</td>
</tr>
<tr>
<td>زرافة</td>
<td>zarāfa</td>
<td></td>
<td>zürafa</td>
<td>sucre</td>
<td>sucre</td>
<td>Giraffe</td>
<td>giraffa</td>
<td>giraffe</td>
</tr>
<tr>
<td>سكر</td>
<td>sukkar</td>
<td></td>
<td>şeker</td>
<td>coton</td>
<td>coton</td>
<td>Zucker</td>
<td>zucchero</td>
<td>sugar</td>
</tr>
<tr>
<td>قطن</td>
<td>quţn</td>
<td></td>
<td>alkol</td>
<td>coton</td>
<td></td>
<td>Alkohol</td>
<td>cotone</td>
<td>cotton</td>
</tr>
<tr>
<td>الكحول</td>
<td>al-kuḥūl</td>
<td>alcohol</td>
<td>al-kuḥūl</td>
<td>sucre</td>
<td></td>
<td>Sirup</td>
<td>alcol</td>
<td>alcohol</td>
</tr>
<tr>
<td>شراب</td>
<td>šarāb</td>
<td></td>
<td>al-jabr</td>
<td>coton</td>
<td></td>
<td>Algebra</td>
<td>sciroppo</td>
<td>syrup</td>
</tr>
<tr>
<td>الجبر</td>
<td>al-jabr</td>
<td></td>
<td>gazæl</td>
<td>coton</td>
<td></td>
<td></td>
<td>algebra</td>
<td>gazæl</td>
</tr>
<tr>
<td>غزال</td>
<td>ġazāl</td>
<td></td>
<td>gazæl</td>
<td>sirop</td>
<td></td>
<td></td>
<td>gazzella</td>
<td>gazelle</td>
</tr>
</tbody>
</table>

Mapping lexicons across languages.
Cross-lingual lexicon induction
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

**Borrowing**
- Content words of foreign origin, assimilated in the language and aren’t perceived as foreign

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

<table>
<thead>
<tr>
<th>Language</th>
<th>Example</th>
<th>Language</th>
<th>Example</th>
<th>Language</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>New York</td>
<td>Arabic</td>
<td>sukkar</td>
<td>Latin</td>
<td>nocte</td>
</tr>
<tr>
<td>Yoruba</td>
<td>Niu Yoki</td>
<td>*transliterated</td>
<td>zuccarum</td>
<td>French</td>
<td>nuit</td>
</tr>
<tr>
<td>Russian</td>
<td>Нью-Йорк</td>
<td>French</td>
<td>sucre</td>
<td>Spanish</td>
<td>noche</td>
</tr>
<tr>
<td>Arabic</td>
<td>نيويورك</td>
<td>German</td>
<td>Zucker</td>
<td>Italian</td>
<td>notte</td>
</tr>
<tr>
<td>Hebrew</td>
<td>ניו יורק</td>
<td>Italian</td>
<td>zucchero</td>
<td>Portuguese</td>
<td>noite</td>
</tr>
</tbody>
</table>
Transliteration models

- FSTs Knight & Graehl ‘98
- Noisy channel approaches Al-Onaizan & Knight ‘02, Virga & Khudanpur ‘03
- String similarity and temporal similarity of distributions in comparable corpora Klementiev & Roth ‘06
- Phonetic similarity and temporal similarity of distributions Tao et al. ‘06
- Decipherment approaches to phonetic mapping in non-parallel corpora Ravi & Knight ‘09
- CRFs Ganesh et al.’08, Ammar et al. ‘12
Transliteration

- LSTMs with attention *Rosca & Breuel’16*
- Exact Hard Monotonic Attention for Character-Level Transduction *Wu & Cotterell’19*

![Diagram showing examples of source and target strings for each task. Tag guides transduction in morphological inflection.](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)
- 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**

- Peripheral
- Assimilated
- Core

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

**Cognates**

- Peripheral
- Assimilated
- Core

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

<table>
<thead>
<tr>
<th>Arabic *transliterated</th>
<th>سكر</th>
<th>sukkar</th>
<th>Latin</th>
<th>nocte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td></td>
<td>zuccarum</td>
<td>French</td>
<td>nuit</td>
</tr>
<tr>
<td>German</td>
<td></td>
<td>Zucker</td>
<td>Spanish</td>
<td>noche</td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td>zucchero</td>
<td>Italian</td>
<td>noite</td>
</tr>
<tr>
<td>English</td>
<td></td>
<td>sugar</td>
<td>Portuguese</td>
<td>noapte</td>
</tr>
</tbody>
</table>
## Arabic--Swahili borrowing examples

<table>
<thead>
<tr>
<th>English</th>
<th>Arabic (Semitic)</th>
<th>Swahili (Bantu)</th>
<th>Phonological &amp; morphological integration</th>
</tr>
</thead>
</table>
| fever   | حممي (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 CNNs 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

- Bilingual embeddings
- Multilingual embeddings
- Subword-based multilingual embeddings
- Subword-based multilingual embeddings with incorporated morphological and phonological knowledge
- Bilingual lexicon induction via embedding similarity

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

- Pick a language that you speak

- Read about the history of this language, 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?