Constraint-Based Models of Lexical Borrowing

Yulia Tsvetkov Waleed Ammar Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{ytsvetko, wammar, cdyer}@cs.cmu.edu

Abstract

Linguistic borrowing is the phenomenon of transferring linguistic constructions (lexical, phonological, morphological, and syntactic) from a "donor" language to a "recipient" language as a result of contacts between communities speaking different languages. Borrowed words are found in all languages, and—in contrast to cognate relationships—borrowing relationships may exist across unrelated languages (for example, about 40% of Swahili's vocabulary is borrowed from Arabic). In this paper, we develop a model of morpho-phonological transformations across languages with features based on universal constraints from Optimality Theory (OT). Compared to several standard—but linguistically naïve—baselines, our OT-inspired model obtains good performance with only a few dozen training examples, making this a cost-effective strategy for sharing lexical information across languages.

1 Introduction

We may imagine that globalization is a modern phenomenon, but the lexicons of the world's languages attest to the fact that robust interaction between communities of speakers of different languages is widespread throughout history. Language contact breeds *linguistic borrowing*—a phenomenon as old as language itself—adoption and nativization of phonemes, morphemes, words, and syntactic constructions from another language (Thomason and Kaufman, 2001).

Contact-induced borrowing is a fundamental research topic in linguistics; however, in computational linguistics, very limited work has addressed modeling this phenomenon. The problem we address is the identification of plausible donor words (in the donor language) given a loanword (in the recipient language), and vice versa, identification of loanwords given a donor. For example, given a Swahili loanword *safari* 'journey', our model identifies its Arabic donor سفريه (*sfryh*)[1] 'journey' (§2). Although at a high level, this is an instance of the well-known problem of modeling string transductions, our interest is being able to identify correspondences across languages with minimal supervision, so as to make the technique applicable in low-resource settings. To reduce the supervision burden, we propose a model that includes awareness of the morpho-phonological repair strategies that native speakers of a language subconsciously employ to adapt a loanword to phonological constraints of the recipient language (§3). To this end, we use constraint-based theories of phonology, as exemplified by Optimality Theory (OT) (Prince and Smolensky, 2008; McCarthy, 2009), which non-computational linguistic work has demonstrated to be particularly well suited to account for phonologically complex borrowing processes (Kang, 2011). We operationalize OT constraints as features in our borrowing model (§4).

We conduct a case study on Arabic and Swahili, two unrelated languages with a long history of contact; we then apply the model to additional language pairs (§5). The proposed approach significantly outperforms transliteration and cognate discovery models (§6).

[1] We use Buckwalter notation to write Arabic glosses.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 598–608,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

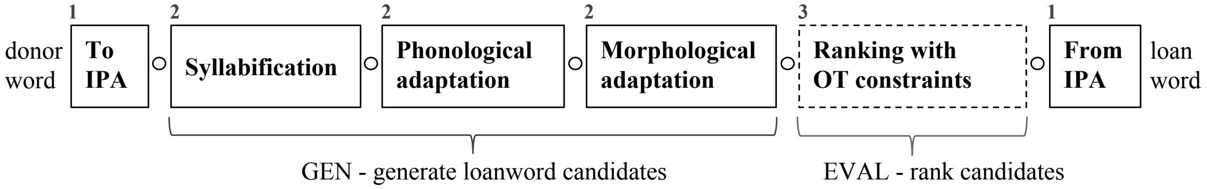

Figure 1: Our morpho-phonological borrowing model conceptually has three main parts: (1) conversion of orthographic word forms to pronunciations in IPA format; (2) generation of loanword pronunciation candidates; (3) ranking of generated candidates using Optimality-Theoretic constraints. Part (1) and (2) are rule-based, (1) uses pronunciation dictionaries, (2) is based on prior linguistic studies; part (3) is learned. In (3) we learn OT constraint weights from a few dozen automatically extracted training examples.

2 Methodology

Our task is to identify plausible donor–loan word pairs in a language pair. While modeling string transductions is a well-studied problem in NLP, we wish to be able to learn the cross-lingual correspondences from minimal amounts of data, so we propose a linguistically-motivated approach: we formulate a scoring model inspired by Optimality Theory (OT; discussed below), in which borrowing candidates are ranked by universal constraints posited to underly the human faculty of language, and the candidates are determined by transduction processes articulated in prior studies of contact linguistics.

As shown in figure 1, our model is conceptually divided into three main parts: (1) mapping of orthographic word forms in two languages into a common space of their phonetic representation; (2) generation of loanword pronunciation candidates from a donor word; (3) ranking of generated loanword candidates, based on linguistic constraints of the donor and recipient languages. Parts (1) and (2) are rule-based; whereas (3) is learned. Each component of the model is discussed in detail in the following sections.

The model is implemented within a finite-state cascade. Parts (1) and (2) amount to unweighted string transformation operations. In (1), we convert orthographic word forms to their pronunciations in the International Phonetic Alphabet (IPA), these are pronunciation transducers. In (2) we syllabify donor pronunciations, then perform insertion, deletion, and substitution of phonemes and morphemes (affixes), to generate multiple loanword candidates from a donor word. Although string transformation transducers in (2) can generate loanword candidates that are not found in a recipient language vocabulary, such candidates are filtered out due to composition with the recipient language lexicon acceptor.

We perform string transformations from donor to recipient (recapitulating the historical process). However, the resulting relation (i.e., the final composed transducer) is a bidirectional model which can just as well be used to reason about underlying donor forms given recipient forms. To employ the model in a specific direction, one needs to optimize parameters—weights on transitions—to generate a desired set of outputs from a specific input. Our model is trained to discriminate a donor word given a loanword. In part (3), candidates are "evaluated" (i.e., scored) with a weighted sum of universal constraint violations. The non-negative weights, which we call "cost vector", constitute our model parameters and are learned using a small training set of donor–recepient pairs. We use a shortest path algorithm to find the path with the minimal cost.

OT: constraint-based evaluation Our decision to evaluate borrowing candidates by weighting counts of "constraint violations" is based on Optimality Theory, which has shown that complex surface phenomena can be well-explained as the interaction of constraints on the form of outputs and the relationships of inputs and outputs (Kager, 1999). Although our linear scoring scheme departs from OT's standard evaluation assumptions (namely, the assumption of an ordinal constraint ranking and strict dominance rather than constraint "weighting"), we are still able to obtain effective models.

Although originally a theory of monolingual phonology, OT has been adapted to account for borrowing by treating the donor language word as the underlying form for the recipient language; that is, the phonological system of the recipient language is encoded as a system of constraints, and these constraints account for how the donor word is adapted when borrowed. There has been substantial prior work in linguistics on borrowing in the OT paradigm (Yip, 1993; Davidson and Noyer, 1997; Jacobs and Gussenhoven, 2000; Kang, 2003; Broselow, 2004; Adler, 2006; Rose and Demuth, 2006; Kenstowicz and Suchato, 2006; Kenstowicz, 2007; Mwita, 2009), but none of it has led to computational realizations.

3 Generating loanword candidates

In this section, we use the Arabic–Swahili language-pair to describe the prototypical linguistic adaptation processes that words undergo when borrowed. Then, we describe how we model these processes.[2]

3.1 Case study: Arabic–Swahili borrowing

The Swahili lexicon has been influenced by Arabic due to a prolonged period of language contact in the Indian Ocean trading (800 A.D.–1920), as well as the influence of Islam (Rothman, 2002). According to several independent studies, Arabic loanwords constitute from 18% (Hurskainen, 2004) to 40% (Johnson, 1939) of Swahili word types.

Despite a strong susceptibility of Swahili to borrowing and a large fraction of Swahili words originating from Arabic, the two languages are typologically distinct with profoundly dissimilar phonological and morpho-syntactic systems. Therefore, Arabic loanwords have been substantially adapted to conform to Swahili phonotactics, which we survey briefly. First, Arabic has five syllable patterns:[3] CV, CVV, CVC, CVCC, and CVVC (McCarthy, 1985, pp. 23–28), whereas Swahili (like other Bantu languages) is characterized by the syllable ending with a vowel and CV or V syllable structure. At the segment level, Swahili loanword adaptation thus involves extensive vowel epenthesis in consonant clusters and at a syllable final position if the syllable ends with a consonant, e.g., : كتاب (*ktAb*) → *kitabu* 'book' (Polomé, 1967; Schadeberg, 2009; Mwita, 2009). Second, phonological adaptation in Swahili loanwords includes shortening of vowels (unlike Arabic, Swahili does not have phonemic length); substitution of consonants that are found in Arabic but not in Swahili (e.g., emphatic (pharyngealized) /tˤ/→/t/, voiceless velar fricative /x/→/k/, dental fricatives /θ/→/s/, /ð/→/z/, and the voiced velar fricative /ɣ/→/g/); adoption of Arabic phonemes that were not originally present in Swahili /θ/, /ð/, /ɣ/ (e.g., تحذير (*tH*yr*)→ *tahadhari* 'warning'); degemination of Arabic geminate consonants (e.g., شرّ (*$r~*)→ *shari* 'evil'). Finally, adapted loanwords can freely undergo Swahili inflectional and derivational processes, e.g., الوزير (*Alwzyr*) → *waziri* 'minister', *mawaziri* 'ministers', *kiuwaziri* 'ministerial' (Zawawi, 1979; Schadeberg, 2009).

3.2 Arabic–Swahili borrowing transducers

We describe unweighted transducers for pronunciation, syllabification, and morphological and phonological adaptation. An example that illustrates some of the possible string transformations by individual components of the model is shown in figure 2. The goal of these transducers is to minimally overgenerate Swahili adapted forms of Arabic words, based on the adaptations described above.

Pronunciation. Based on the IPA, we assign shared symbols to sounds that exist in both sound systems of Arabic and Swahili (e.g., nasals /n/, /m/; voiced stops /b/, /d/), and language-specific unique symbols to sounds that are unique to the phonemic inventory of Arabic (e.g., pharyngeal voiced and voiceless fricatives /ħ/, /ʕ/) or Swahili (e.g., velar nasal /ŋ/). For Swahili, we construct a pronunciation dictionary based on the Omniglot grapheme-to-IPA mapping.[4] In Arabic, we use the CMU Arabic vowelized pronunciation dictionary containing about 700K types which has an average of four pronunciations per word (Metze et al., 2010).[5] We then design four transducers—Arabic and Swahili word-to-IPA and IPA-to-word transducers—each as a union of linear chain transducers, as well as one acceptor per pronunciation dictionary listing.

[2] For simplicity, we subsume Omani Arabic and other historical dialects of Arabic under the label "Arabic"; similarly, we subsume Swahili, its dialects and protolanguages under "Swahili".

[3] C stands for consonant, and V for vowel.

[4] www.omniglot.com

[5] Since we are working at the level of word types which have no context, we cannot disambiguate the intended form, so we include all options. For example, for the input word كتابا (*ktAbA*) 'book.sg.indef', we use both pronunciations /kitɑbɑ/ and /kuttɑbɑ/.

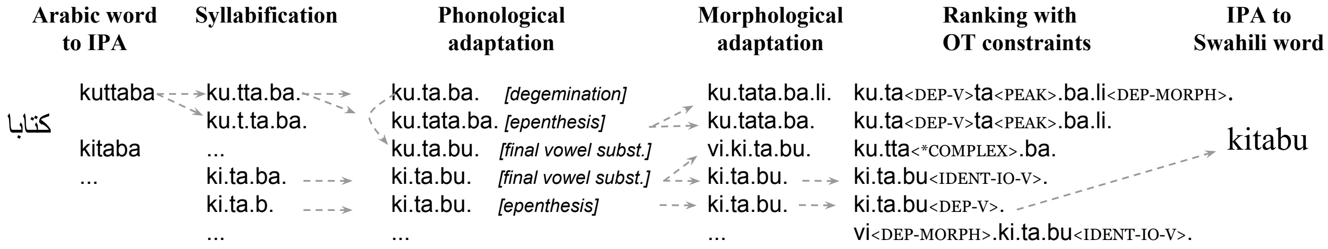

Figure 2: An example of an Arabic word كتابا (*ktAbA*) 'book.sg.indef' transformed by our model into a Swahili loanword *kitabu*.

Syllabification. Arabic words borrowed into Swahili undergo a repair of violations of the Swahili segmental and phonotactic constraints, for example via vowel epenthesis in a consonant cluster. Importantly, *repair depends upon syllabification*. To simulate plausible phonological repair processes, we generate multiple syllabification variants for input pronunciations. The syllabification transducer optionally inserts syllable separators between phones. For example, for an input phonetic sequence /kuttɑbɑ/, the output strings include /ku.t.tɑ.bɑ/, /kut.tɑ.bɑ/, and /ku.ttɑ.bɑ/ as syllabification variants; each variant violates different constraints and consequently triggers different phonological adaptation.

Phonological adaptation. Phonological adaptation of syllabified phone sequences is the crux of the loanword adaptation process. We implement phonological adaptation transducers as a composition of plausible context-dependent insertions, deletions, and substitutions of phone subsets, based on prior studies summarized in §3.1. In what follows, we list phonological adaptation components in the order of transducer composition in the borrowing model. The **vowel deletion** transducer shortens Arabic long vowels and vowel clusters. The **consonant degemination** transducer shortens Arabic geminate consonants, e.g., it degeminates /tt/ in /ku.ttɑ.bɑ/, outputting /ku.tɑ.bɑ/. The **substitution of similar phonemes** transducer substitutes similar phonemes and phonemes that are found in Arabic but not in Swahili (Polomé, 1967, p. 45). For example, the emphatic /t^{ʕ}/, /d^{ʕ}/, /s^{ʕ}/ are replaced by the corresponding non-emphatic segments [t], [d], [s]. The **vowel epenthesis** transducer inserts a vowel between pairs of consonants (/ku.ttɑ.bɑ/ → /ku.tatɑ.bɑ/), and at the end of a syllable, if the syllable ends with a consonant (/ku.t.tɑ.bɑ/ → /ku.tɑ.tɑ.bɑ/). Sometimes it is possible to predict the final vowel of a word, depending on the word-final coda consonant of its Arabic counterpart: /u/ or /o/ added if an Arabic donor ends with a labial, and /i/ or /e/ added after coronals and dorsals (Mwita, 2009). Following these rules, the **final vowel substitution** transducer complements the inventory of final vowels in loanword candidates.

Morphological adaptation. Both Arabic and Swahili have significant morphological processes that alter the appearance of lemmas. To deal with morphological variants, we construct morphological adaptation transducers that optionally strip Arabic concatenative affixes and clitics, and then optionally append Swahili affixes, generating a superset of all possible loanword hypotheses. We obtain the list of Arabic affixes from the Arabic morphological analyzer SAMA (Maamouri et al., 2010); the Swahili affixes are taken from a hand-crafted Swahili morphological analyzer (Littell et al., 2014).

4 Learning constraint weights

Due to the computational problems of working with OT (Eisner, 1997; Eisner, 2002), we make simplifying assumptions by (1) bounding the theoretically infinite set of underlying forms with a small linguistically-motivated subset of allowed transformations on donor pronunciations, as described in §3; (2) imposing a priori restrictions on the set of the surface realizations by intersecting the candidate set with the recipient pronunciation lexicon; (3) assuming that the set of constraints is finite and regular (Ellison, 1994); and (4) assigning linear weights to constraints, rather than learning an ordinal constraint ranking (Boersma and Hayes, 2001; Goldwater and

Faithfulness constraints	
MAX-IO-MORPH	no (donor) affix deletion
MAX-IO-C	no consonant deletion
MAX-IO-V	no vowel deletion
DEP-IO-MORPH	no (recipient) affix epenthesis
DEP-IO-V	no vowel epenthesis
IDENT-IO-C	no consonant substitution
IDENT-IO-C-M	no subst. in manner of pronunciation
IDENT-IO-C-A	no subst. in place of articulation
IDENT-IO-C-S	no subst. in sonority
IDENT-IO-C-P	no pharyngeal consonant substitution
IDENT-IO-C-G	no glottal consonant substitution
IDENT-IO-C-E	no emphatic consonant substitution
IDENT-IO-V	no vowel substitution
IDENT-IO-V-O	no subst. in vowel openness
IDENT-IO-V-R	no subst. in vowel roundness
IDENT-IO-V-F	no subst. in vowel frontness
IDENT-IO-V-FIN	no final vowel substitution

Table 1: Faithfulness constraints prefer pronounced realizations completely congruent with their underlying forms.

Markedness constraints	
NO-CODA	syllables must not have a coda
ONSET	syllables must have onsets
PEAK	there is only one syllabic peak
SSP	complex onsets rise in sonority, complex codas fall in sonority
*COMPLEX-S	no consonant clusters on syllable margins
*COMPLEX-C	no consonant clusters within a syllable
*COMPLEX-V	no vowel clusters

Table 2: Markedness constraints impose language-specific structural well-formedness of surface realizations.

Johnson, 2003).

OT distinguishes "markedness" constraints (McCarthy and Prince, 1995), which detect dispreferred phonetic patterns in the language, and "faithfulness" constraints (Prince and Smolensky, 2008), which ensure correspondences between the underlying form and the surface candidates.[6] The implemented constraints are listed in tables 1 and 2. Faithfulness constraints are integrated in phonological transformation components as transitions following each insertion, deletion, or substitution. Markedness constraints are implemented as standalone identity transducers: inputs are equal outputs, but path weights representing candidate evaluation with respect to violated constraints are different.

The final "loanword transducer" is the composition of all transducers described in §3 and OT constraint transducers. A path in the transducer represents a syllabified phonemic sequence along with (weighted) OT constraints it violates, and shortest path outputs are those, whose cumulative weight of violated constraints is minimal.

OT constraints are realized as features in our linear model, and feature weights are learned in a discriminative training to maximize the accuracy obtained by the loanword transducer on a small development set of donor–recipient pairs. For parameter estimation, we employ the Nelder–Mead algorithm (Nelder and Mead, 1965), a heuristic derivative-free method that iteratively optimizes, based on an objective function evaluation, the convex hull of $n+1$ simplex vertices.[7] The objective function is the "soft accuracy" of the development set, defined as the proportion of correctly identified donor words in the total set of 1-best outputs.

5 Adapting the model to a new language

Although we conduct a thorough case study on the Arabic–Swahili language pair, our methodology can easily be generalized to other language pairs. String transformation operations, as well as OT constraints are language-universal. The only adaptation required is a linguistic analysis to identify plausible morpho-phonological repair strategies for the new language pair (i.e., a subset of allowed insertions, deletions and substitutions of phonemes and morphemes). Since we need only to overgenerate candidates (the OT constraints will filter bad outputs), the effort is minimal relative to many other grammar engineering exercises. The second language-specific component is the grapheme-to-IPA converter. While this can be a non-

[6] To clarify the distinction between faithfulness and markedness constraint groups to the NLP readership, we can draw the following analogy to the components of machine translation or speech recognition: faithfulness constraints are analogical to the translation model or acoustic model (reflecting input), while markedness constraints are analogical to the language model (requiring well-formedness of the output). Without faithfulness constraints, the optimal surface form could differ arbitrarily from the underlying form.

[7] The decision to use Nelder–Mead rather than more conventional gradient-based optimization algorithms was motivated by practical limitations of the finite-state toolkit we used that made computing derivatives with latent structure impractical.

trivial problem in some cases, the problem is well studied, and many under-resourced languages have "phonographic" systems where orthography corresponds to phonology, rather than organically evolved written forms, which makes the mapping problem trivial.

To illustrate the ease with which a language pair can be engineered, we applied our borrowing model to the French–Romanian language pair. Although French and Romanian are sister languages (both descending from Latin), about 12% of Romanian types are true French borrowings that came into the language in the past few centuries (Schulte, 2009). We employ the GLOBALPHONE pronunciation dictionary for French (Schultz and Schlippe, 2014) (we convert it to IPA), and automatically construct a Romanian pronunciation dictionary using Omniglot grapheme-to-IPA conversion rules.

6 Experiments

Our experimental setup is defined as follows. The input to the borrowing model is a loanword candidate in Swahili/Romanian,[8] the outputs are plausible donor words in the Arabic/French monolingual lexicon (i.e., any word in pronunciation dictionary). We train the borrowing model using a small set of training examples, and then evaluate it using a held-out test set. In the rest of this section we describe in detail our datasets, tools, and experimental results.

Resources We employ Arabic–English and Swahili–English bitexts to extract a training set (corpora of sizes 5.4M and 14K sentence pairs, respectively), using a cognate discovery technique (Kondrak, 2001). Phonetically and semantically similar strings are classified as cognates; phonetic similarity is the string similarity between phonetic representations, and semantic similarly is approximated by translation.[9] We thereby extract Arabic and Swahili pairs $\langle a, s \rangle$ that are phonetically similar ($\frac{\Delta(a,s)}{\min(|a|,|s|)} < 0.5$) where $\Delta(a, s)$ is the Levenshtein distance between a and s and that are aligned to the same English word e. FastAlign (Dyer et al., 2013) is used for word alignments. Given an extracted word pair $\langle a, s \rangle$, we also extract word pairs $\{\langle a', s \rangle\}$ for all proper Arabic words a' which share the same lemma with a producing on average 33 Arabic types per Swahili type. We use MADA (Habash et al., 2009) for Arabic morphological expansion.

From the resulting dataset of 490 extracted Arabic–Swahili borrowing examples,[10] we set aside randomly sampled 73 examples (15%) for evaluation,[11] and use the remaining 417 examples for model parameter optimization. For French–Romanian language pair, we use an existing small annotated set of borrowing examples,[12] with 282 training and 50 (15%) randomly sampled test examples.

We use `pyfst`—a Python interface to OpenFst (Allauzen et al., 2007)—for the borrowing model implementation.[13]

Baselines We compare our model to several baselines. In the Levenshtein (L) distance baselines we chose the closest word (either surface or pronunciation-based). In the Levenshtein-weighted (L-W) baselines, we evaluate a variant of the Levenshtein distance tuned to identify cognates (Mann and Yarowsky, 2001; Kondrak and Sherif, 2006); this method was identified by Kondrak and Sherif (2006) among the top three cognate identification methods. In the CRF baselines we generate plausible "transliterations" of the input Swahili (or Romanian) words in the donor lexicon using the model of Ammar et al. (2012), with multiple references in a lattice and without reranking. The CRF transliteration model is a linear-chain CRF where we label each source character with a sequence of target characters. The features are label unigrams, label bigrams, and label

[8] Our model does not provide a mechanism for identifying loanwords in the recipient language; we only model the borrowing process. Classifying loanwords in the recipient language is an interesting but ultimately different problem: the ontological status of words in a lexicon is a difficult problem, even for human experts, however, knowledge of cross-lingual correspondences is a valuable feature, and as such, our work can be understood as enabling this.

[9] This cognate discovery technique is sufficient to extract a small training set, but is not generally applicable, as it requires parallel corpora or manually constructed dictionaries to measure semantic similarity. Large parallel corpora are unavailable for most language pairs, including Swahili–English.

[10] In each training/test example one Swahili word corresponds to all extracted Arabic donor words.

[11] We manually verified that our test set contains clear Arabic–Swahili borrowings. For example, we extract Swahili *kusafiri, safari* and Arabic السفر, يسفر, سفر (*Alsfr, ysAfr, sfr*) all aligned to 'travel'.

[12] `http://wold.clld.org/vocabulary/8`

[13] `https://github.com/vchahun/pyfst`

conjoined with a moving window of source characters. In the OT-uniform baseline, we evaluate the accuracy of the borrowing model with uniform weights, thus shortest paths in the loanwords transducer will be forms violating the fewest constraints.

Evaluation In addition to predictive accuracy on all models (if a model produces multiple hypotheses with the same 1-best weight, we count the proportion of correct outputs in this set), we evaluate two particular aspects of our proposed model: (1) appropriateness of the model family, and (2) the quality of the learned OT constraint weights. The first aspect is designed to evaluate whether the morpho-phonological transformations implemented in the model are required *and* sufficient to generate loanwords from the donor inputs. We report two evaluation measures: model *reachability* and *ambiguity*. Reachability is a percentage of test samples that are reachable (i.e., there is a path from the input test example to a correct output) in the loanword transducer. A naïve model which generates all possible strings would score 100% reachability, but it will be hard to set the model parameters such that it discriminates between good and bad candidates. In order to capture this trade-off, we also report the inherent *ambiguity* of our model, which is the average number of outputs potentially generated per input. A generic Arabic–Swahili transducer, for example, has an ambiguity of 786,998—the size of the Arabic pronunciation lexicon.

Results The borrowing model reachability and ambiguity are listed in table 3. The model obtains high reachability, while significantly reducing the average number of possible outputs per input: in Arabic from 787K to 857 words, in French from 62K to 12. This result shows that the loanword transducer design, based on the prior linguistic analysis, is a plausible model of word borrowing. Yet, there are on average 33 correct Arabic words out of the possible 857 outputs, thus the second part of the model—OT constraint weights optimization—is crucial.

The accuracy results in table 4 show how challenging the task of modeling lexical borrowing between two distinct languages is, and importantly, that orthographic and phonetic baselines including the state-of-the-art generative model of transliteration are not suitable for this task. Phonetic baselines for Arabic–Swahili perform better than orthographic ones, but substantially worse than OT-based models, even if OT constraints are not weighted. Crucially, the performance of the borrowing model with the learned OT weights corroborates the assumption made in numerous linguistic accounts that OT is an adequate analysis of the lexical borrowing phenomenon.

	AR–SW	FR–RO
Reachability	87.7%	82.0%
Ambiguity	857	12

Table 3: The evaluation of the borrowing model design. Reachability is a percentage of donor–recipient pairs that are reachable from a donor to a recipient language. Ambiguity is an average number of outputs that the model generates per one input.

		Accuracy (%)	
		AR–SW	FR–RO
Orthographic	L	8.9	38.0
	CRF	16.4	36.0
Phonetic	L	19.8	26.3
	L-W	19.7	30.7
OT	OT-U	29.3	58.5
	OT	**48.4**	**75.6**

Table 4: The evaluation of the borrowing model accuracy. The baselines are orthographic (surface) and phonetic (based on pronunciation lexicon) Levenshtein distance (L), heuristic Levenshtein distance with lower penalty on vowel updates and similar letter/phone substitutions (L-W), CRF transliteration, and our model with uniform (OT-U) and learned OT constraint weights assignment.

Qualitative evaluation The constraint ranking learned by the borrowing model (constraints are listed in tables 1, 2) is in line with prior linguistic analysis. Space precludes a thorough discussion, but we highlight a few points. In Swahili NO-CODA dominates all other markedness constraints, as expected. Both *COMPLEX-S and *COMPLEX-C, restricting consonant clusters, dominate *COMPLEX-V, confirming that Swahili is more permissive to vowel clusters. SSP—sonority-based constraint—captures a common pattern of consonant clustering, found across languages, and is also learned by our model as undominated by most competitors in Swahili, and as a dominating markedness constraint in Romanian. Finally, vowel epenthesis DEP-IO-V is the most common strategy in Arabic loanword adaptation, and is ranked lower according to the model; however, it is ranked

EN	AR gloss	AR pronunciation	SW syllabification	Violated OT constraints
book	ktAb	kitAb	ki.ta.bu.	IDENT-IO-C-G$\langle A, a\rangle$, DEP-IO-V$\langle \epsilon, u\rangle$
palace	AlqSr	AlqaSr	ka.sri	MAX-IO-MORPH$\langle Al, \epsilon\rangle$, IDENT-IO-C-S$\langle q, k\rangle$, IDENT-IO-C-E$\langle S, s\rangle$, *COMPLEX-C$\langle sr\rangle$, DEP-IO-V$\langle \epsilon, i\rangle$
wage	Ajrh	Aujrah	u.ji.ra.	MAX-IO-V$\langle A, \epsilon\rangle$, ONSET$\langle u\rangle$, DEP-IO-V$\langle \epsilon, i\rangle$, MAX-IO-C$\langle h, \epsilon\rangle$

Table 5: Examples of syllabification and OT constraint violations produced by our borrowing model.

highly in the French–Romanian model, where vowel insertion is rare.

A second interesting by-product of our model is an inferred syllabification. While we did not conduct a systematic quantitative evaluation, higher-ranked Swahili outputs tend to contain linguistically plausible syllabifications, although the syllabification transducer inserts optional syllable boundaries between every pair of phones. This result further attests to the plausible constraint ranking learned by the model. Example Swahili syllabifications[14] along with the OT constraint violations produced by the borrowing model are depicted in table 5.

7 Discussion

The task of modeling borrowing is unexplored in computational linguistics. In this section we first situate the task with respect to two most closely related research directions: modeling transliteration and cognate forms. We then motivate the new line of research proposed in this work: modeling borrowing.

Borrowing vs. transliteration Borrowing is not transliteration. Transliteration refers to writing in a different *orthography*, whereas borrowing refers to *expanding a language* to include words adapted from another language. Unlike borrowing, transliteration is more amenable to orthographic—rather than morpho-phonological—features, although see (Knight and Graehl, 1998). Borrowed words might have begun as transliterations, but a characteristic of borrowed words is that they become assimilated in the linguistic system of the recipient language, and became regular content words, e.g., 'orange' and 'sugar' are English words borrowed from Arabic نارنج (*nArnj*) and السكر (*Alskr*), respectively.

Borrowing vs. inheritance Cognates are words in related languages inherited from one word in a common ancestral language (the proto-language). Loanwords, on the other hand, can occur between any languages, either related or not, that historically came into contact. Theoretical analysis of cognates has tended to be concerned with a diachronic point of view, i.e., modeling word changes across time. While of immense scientific interest, language processing applications are arguably better served by models of synchronic processes, peculiar to loanword analysis.

Why borrowing? Borrowing is a distinctive and pervasive phenomenon: *all* languages borrowed from other languages at some point in their lifetime, and borrowed words constitute a large fraction of most language lexicons. Another important property of borrowing is that in adaptation of borrowed items, changes in words are systematic, knowledge of morphological and phonological patterns in a language can be used to predict how borrowings will be realized in that language, without having to list them all. Therefore, modeling of borrowing is a task well-suited for computational approaches.

Our suggestion in this work is that we can identify borrowing relations between resource-limited languages and resource-rich donor languages, such as English, French, Spanish, Arabic, Chinese, and Russian. For example, 30–70% of the vocabulary in Vietnamese, Cantonese, and Thai—relatively resource-limited languages spoken by hundreds of millions of people—are borrowed from Chinese and English. Similarly, African languages have been greatly influenced by Arabic, Spanish, English, and French—widely spoken languages such as Swahili, Zulu, Malagasy, Hausa, Tarifit, Yoruba contain up to 40% of loanwords. Indo-Iranian languages—Hindustani, Hindi, Urdu, Bengali, Persian, Pashto—spoken by 860 million, also extensively borrowed from Arabic and English (Haspelmath and Tadmor, 2009). In

[14] We chose examples from the Arabic–Swahili system because this is a more challenging case due to linguistic discrepancies.

short, at least a billion people are speaking resource-scarce languages whose lexicons are heavily borrowed from resource-rich languages.

Why is this important? Lexical translations or alignments extracted from large parallel corpora have been widely used to project annotations from high- to low-resource languages (Hwa et al., 2005; Täckström et al., 2013; Ganchev et al., 2009; Tsvetkov et al., 2014, *inter alia*). Unfortunately, parallel resources are unavailable for the majority of resource-limited languages. Loanwords can be used as a source of cross-lingual links complementary to lexical alignments. This holds promise for applying existing cross-lingual methods and bootstrapping linguistic resources in languages where no parallel data is available.

8 Related work

With the exception of a study conducted by Blair and Ingram (2003) on generation of borrowed phonemes in English–Japanese language pair (the method does not generalize from borrowed phonemes to borrowed words, and does not rely on linguistic insights), we are not aware of any prior work on computational modeling of lexical borrowing. Few papers only mention or tangentially address borrowing, we briefly list them here. Daumé III (2009) focuses on areal effects on linguistic typology, a broader phenomenon that includes borrowing and genetic relations across languages. This study is aimed at discovering language areas based on typological features of languages. Garley and Hockenmaier (2012) train a maxent classifier with character n-gram and morphological features to identify anglicisms (which they compare to loanwords) in an online community of German hip hop fans. List and Moran (2013) have published a toolkit for computational tasks in historical linguistics but remark that "Automatic approaches for borrowing detection are still in their infancy in historical linguistics."

Two related lines of research are transliteration and cognate identification. Knight and Graehl (1998), Al-Onaizan and Knight (2002) developed a finite-state generative model of transliteration, and successfully applied it to Arabic–English named entity translation. Mann and Yarowsky (2001) and Kondrak (2001) identify cognate pairs, based on the learned surface and phonetic similarities, respectively. As our experiments confirm, orthographic and phonetic transliteration and string edit distance methods are not adequate models for the complex borrowing phenomena.

9 Conclusion

Given a loanword, our model identifies plausible donor words in a contact language. We show that a discriminative model with Optimality Theoretic features effectively models systematic phonological changes in Arabic–Swahili loanwords. We also found that the model and methodology is generally applicable to other language pairs with minimal engineering effort.

This paper makes two contributions: (1) To the best of our knowledge, this is the first computational model of lexical borrowing. (2) While there are implementations of OT (Hayes et al., 2013), they are used chiefly to facilitate linguistic analysis.

There are numerous research questions that we would like to explore further. Is it possible to monolingually identify borrowed words in a language? Can we automatically identify a donor language (or its phonological properties) for a borrowed word? Since languages may borrow from many sources, can jointly modeling this process lead to better performance? Can we reduce the amount of language-specific engineering required to deploy our model? Can we integrate knowledge of borrowing in downstream NLP applications? We intend to address these questions in future work.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533. Computational resources were provided by Google in the form of a Google Cloud Computing grant and the NSF through the XSEDE program TG-CCR110017. We are grateful to Nathan Schneider, David Mortensen, Archna Bhatia, Shuly Wintner, Shay Cohen, David Bamman, Noah Smith, and Lori Levin for extensive discussions and constructive feedback.

References

Allison N Adler. 2006. Faithfulness and perception in loanword adaptation: A case study from Hawaiian. *Lingua*, 116(7):1024–1045.

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in Arabic text. In *Proc. of the ACL workshop on Computational Approaches to Semitic Languages*, pages 1–13.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

Waleed Ammar, Chris Dyer, and Noah A. Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proc. of NEWS workshop at ACL*.

Alan D Blair and John Ingram. 2003. Learning to predict the phonological structure of English loanwords in Japanese. *Applied Intelligence*, 19(1-2):101–108.

Paul Boersma and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, 32(1):45–86.

Ellen Broselow. 2004. Language contact phonology: richness of the stimulus, poverty of the base. In *Proc. of NELS*, volume 34, pages 1–22.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proc. of NAACL*, pages 593–601.

Lisa Davidson and Rolf Noyer. 1997. Loan phonology in Huave: nativization and the ranking of faithfulness constraints. In *Proc. of WCCFL*, volume 15, pages 65–79.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. of NAACL*.

Jason Eisner. 1997. Efficient generation in primitive Optimality Theory. In *Proc. of EACL*, pages 313–320.

Jason Eisner. 2002. Comprehension and compilation in Optimality Theory. In *Proc. of ACL*, pages 56–63.

T Mark Ellison. 1994. Phonological derivation in Optimality Theory. In *Proc. of CICLing*, pages 1007–1013.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proc. of ACL*, pages 369–377.

Matt Garley and Julia Hockenmaier. 2012. Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proc. of ACL*, pages 135–139.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proc. of the Stockholm workshop on variation within Optimality Theory*, pages 111–120.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proc. of MEDAR*, pages 102–109.

Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bruce Hayes, Bruce Tesar, and Kie Zuraw. 2013. OTSoft 2.3.2. *software package, http://www.linguistics.ucla.edu/people/hayes/otsoft*.

Arvi Hurskainen. 2004. Loan words in Swahili. In Katrin Bromber and Birgit Smieja, editors, *Globalisation and African Languages*, pages 199–218. Walter de Gruyter.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3).

Haike Jacobs and Carlos Gussenhoven. 2000. Loan phonology: perception, salience, the lexicon and OT. *Optimality Theory: Phonology, syntax, and acquisition*, pages 193–209.

Frederick Johnson. 1939. *Standard Swahili-English dictionary*. Oxford University Press.

René Kager. 1999. *Optimality Theory*. Cambridge University Press.

Yoonjung Kang. 2003. Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology*, 20(2):219–274.

Yoonjung Kang. 2011. Loanword phonology. In Mark van Oostendorp, Colin Ewen, Elizabeth Hume, and Keren Rice, editors, *Companion to Phonology*. Wiley–Blackwell.

Michael Kenstowicz and Atiwong Suchato. 2006. Issues in loanword adaptation: A case study from Thai. *Lingua*, 116(7):921–949.

Michael Kenstowicz. 2007. Salience and similarity in loanword adaptation: a case study from Fijian. *Language Sciences*, 29(2):316–340.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proc. of the Workshop on Linguistic Distances*, pages 43–50.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proc. of NAACL*, pages 1–8.

Johann-Mattis List and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proc. of ACL (System Demonstrations)*, pages 13–18.

Patrick Littell, Kaitlyn Price, and Lori Levin. 2014. Morphological parsing of Swahili using crowdsourced lexical resources. In *Proc. of LREC*.

Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, and Seth Kulick. 2010. LDC Standard Arabic morphological analyzer (SAMA) v. 3.1. *LDC Catalog No. LDC2010L01. ISBN*, pages 1–58563.

Gideon S Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proc. of HLT-NAACL*, pages 1–8.

John J McCarthy and Alan Prince. 1995. Faithfulness and reduplicative identity. *Beckman et al. (Eds.)*, pages 249–384.

John J McCarthy. 1985. *Formal problems in Semitic phonology and morphology*. Ph.D. thesis, MIT.

John J McCarthy. 2009. *Doing Optimality Theory: Applying theory to data*. John Wiley & Sons.

Florian Metze, Roger Hsiao, Qin Jin, Udhyakumar Nallasamy, and Tanja Schultz. 2010. The 2010 CMU GALE speech-to-text system. In *Proc. of INTERSPEECH*, pages 1501–1504.

Leonard Chacha Mwita. 2009. The adaptation of Swahili loanwords from Arabic: A constraint-based analysis. *Journal of Pan African Studies*.

John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *Computer journal*, 7(4):308–313.

Edgar C Polomé. 1967. *Swahili Language Handbook*. ERIC.

Alan Prince and Paul Smolensky. 2008. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Yvan Rose and Katherine Demuth. 2006. Vowel epenthesis in loanword adaptation: Representational and phonetic considerations. *Lingua*, 116(7):1112–1139.

Norman C Rothman. 2002. Indian Ocean trading links: The Swahili experience. *Comparative Civilizations Review*, 46:79–90.

Thilo C Schadeberg. 2009. Loanwords in Swahili. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 76–102. Max Planck Institute for Evolutionary Anthropology.

Kim Schulte. 2009. Loanwords in Romanian. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 230–259. Max Planck Institute for Evolutionary Anthropology.

Tanja Schultz and Tim Schlippe. 2014. GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proc. of LREC*.

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press Edinburgh.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proc. of ACL*, pages 248–258.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Moira Yip. 1993. Cantonese loanword phonology and Optimality Theory. *Journal of East Asian Linguistics*, 2(3):261–291.

Sharifa Zawawi. 1979. *Loan words and their effect on the classification of Swahili nominals*. Leiden: E.J. Brill.