Journal of Artificial Intelligence Research ? (2015) ?-? Submitted 03/15; published ?/15

Cross-Lingual Bridges with Models of Lexical Borrowing

Yulia Tsvetkov YTSVETKO@CS.CMU.EDU
Chris Dyer CDYER@CS.CMU.EDU
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

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 the unrelated language Arabic). In this work, we develop a model of morpho-phonological transformations across languages. Its features are based on universal constraints from Optimality Theory (OT), and we show that compared to several standard—but linguistically more naïve—baselines, our OT-inspired model obtains good performance at predicting donor forms from borrowed forms with only a few dozen training examples, making this a cost-effective strategy for sharing lexical information across languages. We demonstrate applications of the lexical borrowing model in machine translation, using resource-rich donor language to obtain translations of out-of-vocabulary loanwords in a lower resource language. Our framework obtains substantial improvements (up to 1.6 BLEU) over standard baselines.

1. Introduction

State-of-the-art natural language processing (NLP) tools, such as text parsing, speech recognition and synthesis, text and speech translation, semantic analysis and inference, rely on availability of language-specific data resources that exist only for a few resource-rich languages. To make NLP tools available in more languages, techniques have been developed for projecting such resources from resource-rich languages using parallel (translated) data as a bridge for cross-lingual part-of-speech tagging (Yarowsky, Ngai, & Wicentowski, 2001; Das & Petrov, 2011; Li, Graça, & Taskar, 2012; Täckström, Das, Petrov, McDonald, & Nivre, 2013), syntactic parsing (Wu, 1997; Kuhn, 2004; Smith & Smith, 2004; Hwa, Resnik, Weinberg, Cabezas, & Kolak, 2005; Xi & Hwa, 2005; Burkett & Klein, 2008; Snyder, Naseem, & Barzilay, 2009; Ganchev, Gillenwater, & Taskar, 2009; Tiedemann, 2014), word sense tagging (Diab & Resnik, 2002), semantic role labeling (Padó & Lapata, 2009; Kozhevnikov & Titov, 2013), metaphor identification (Tsvetkov, Boytsov, Gershman, Nyberg, & Dyer, 2014), and others. The limiting reagent in these methods is parallel data. While small parallel corpora do exist for many languages (Smith, Saint-Amand, Plamada, Koehn, Callison-Burch, & Lopez, 2013), suitably large parallel corpora are expensive, and these typically exist only for English and a few other geopolitically or economically important language pairs. Furthermore, while English is a high-resource language, it is linguistically a typological

outlier in a number of respects (e.g., relatively simple morphology, complex system of verbal auxiliaries, large lexicon, etc.), and the assumption of construction-level parallelism that projection techniques depend on is thus questionable. Given this state of affairs, there is an urgent need for methods for establishing lexical links across languages that do not rely on large-scale parallel corpora. Without new strategies, most of the 7,000+ languages in the world—many with millions of speakers—will remain resource-poor from the standpoint of NLP.

We advocate a novel approach to automatically constructing language-specific resources, even in languages with no resources other than raw text corpora. Our main motivation is research in **linguistic borrowing**—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 (Thomason & Kaufman, 2001). Borrowed words (also called loanwords, e.g., in Figure 1) are lexical items adopted from another language and integrated (nativized) in the recipient language. Borrowing occurs typically on the part of minority language speakers, from the language of wider communication into the minority language (Sankoff, 2002); that is one reason why donor languages often bridge between resource-rich and resource-limited languages. 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 (10–70%) of most language lexicons (Haspelmath, 2009).

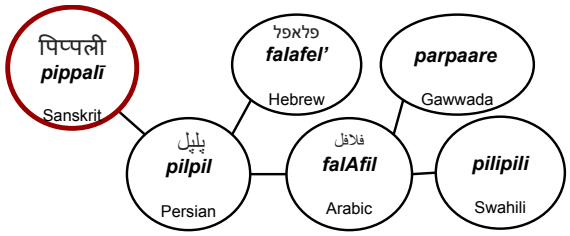

Figure 1: An example of the multilingual borrowing from Sanskrit into typologically diverse, low- and high-resource languages (Haspelmath & Tadmor, 2009).

Loanword nativization is primarily a phonological process. Donor words undergo phonological repairs to adapt a foreign word to the segmental, phonotactic, suprasegmental and morpho-phonological constraints of the recipient language (Holden, 1976; Van Coetsem, 1988; Ahn & Iverson, 2004; Kawahara, 2008; Hock & Joseph, 2009; Calabrese & Wetzels, 2009; Kang, 2011, *inter alia*). Common phonological repair strategies include feature/phoneme epenthesis, elision, degemination, and assimilation. When speakers encounter a foreign word (either a lemma or an inflected form), they analyze it morphologically as a stem, and morphological loanword integration thus amounts to selecting an appropriate donor surface form (out of existing inflections of the same lemma), and applying the recipient language morphology (Repetti, 2006). Adapted loanwords can freely undergo recipient language inflectional and derivational processes. Nouns are borrowed preferentially, then other parts of speech, then affixes, inflections, and phonemes (Whitney, 1881; Moravcsik, 1978; Myers-Scotton, 2002, p. 240).

Although borrowing is pervasive and a topic of enduring interest for historical and theoretical linguists (Haugen, 1950; Weinreich, 1979), only limited work in computational modeling has addressed this phenomenon. However, it is a topic well-suited to computational models (e.g., the systematic phonological changes that occur during borrowing can be modeled using established computational primitives such as finite state transducers), and models

of borrowing have useful applications. Our work can be summarized as the development of a computational model of lexical borrowing and an exploration of its applications to augment language resources and computational approaches to NLP in resource-limited languages. Specifically, we demonstrate how multilingual dictionaries extracted using models of borrowing improve resource-limited statistical machine translation (MT), using a pivoting paradigm where the borrowing pair and the translation pair have only a single language in common.

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. For example, given a Swahili loanword *safari* 'journey', our model identifies its Arabic donor سفريه (*sfryh*)[1] 'journey' (§3). 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, our model 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 (§3.3). To this end, we use constraint-based theories of phonology, as exemplified by Optimality Theory (OT) (Prince & 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 (§3.4). We conduct a case study on Arabic and Swahili, two phylogenetically unrelated languages with a long history of contact; we then apply the model to additional language pairs (§3.5). We then employ models of lexical borrowing to obtain cross-lingual bridges from loanwords in a low-resource language to their donors in a resource-rich language. The donor language is used as pivot to obtain translations via triangulation of out-of-vocabulary loanwords (§4). We conduct translation experiments with three resource-poor setups: Swahili–English pivoting via Arabic, Maltese–English pivoting via Italic, and Romanian–English[2] pivoting via French. In intrinsic evaluation, Arabic–Swahili, Italian–Maltese, and French–Romanian borrowing models significantly outperform transliteration and cognate discovery models (§5.1). We then provide a systematic quantitative and qualitative analysis of contribution of integrated translations, relative to baselines and oracles, and on corpora of varying sizes (§5.2). The proposed pivoting approach yields substantial improvements (up to +1.6 BLEU) in Swahili–Arabic–English translation, moderate improvement (up to +0.8 BLEU) in Maltese–Italian–English translation, and small (+0.2 BLEU) but statistically significant improvements in Romanian–French–English.

Our contributions are twofold.[3] While there have been software implementations of OT (Hayes, Tesar, & Zuraw, 2013), they have been used chiefly to facilitate linguistic analysis; we show how to use OT to formulate a model that can be learned with less supervision than linguistically naïve models. To the best of our knowledge, this is the first

1. We use Buckwalter notation to write Arabic glosses.
2. Romanian is not resource-poor from MT perspective, but in this work we simulate a resource-poor scenario.
3. This article is a thoroughly revised and extended version of Tsvetkov, Ammar, and Dyer (2015), Tsvetkov and Dyer (2015). We provide a more detailed linguistic background of lexical borrowing and OT. We demonstrate results on a new language, Maltese, to emphasize the generality of the method. Additional extensions include a detailed error analysis and a more complete literature survey.

computational model of lexical borrowing used in a downstream NLP task. Second, we show that lexical correspondences induced using this model can project resources—namely, translations—leading to improved performance in a downstream translation system.

2. Motivation

The task of modeling borrowing is under-explored in computational linguistics, although it has both important practical applications and lends itself to modeling with a variety of established computational techniques. In this section we first situate the task with respect to two most closely related research directions: modeling transliteration and modeling 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 & 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. Whatever their historical origins, synchronically, these words are indistinguishable to most speakers from words that have native ancestral forms in the language. Thus, the morpho-phonological processes that must be accounted for in borrowing models are more complex than is required by transliteration models.

Borrowing vs. inheritance. Cognates are words in related languages that are inherited from a single 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. From a modeling perspective, cognates and borrowed words require separate investigation as loanwords are more likely to display marginal phonotactic (and other phonological) patterns than inherited lexical items. 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, and 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-rich donor languages (such as English, French, Spanish, Arabic, Chinese, or Russian) and resource-limited recipient languages. 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, languages for which numerous data resources have been created. 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 & Tadmor, 2009). In 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, *inter alia*). Unfortunately, large-scale 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 obtained from parallel data or bilingual lexicons. This holds promise for applying existing cross-lingual methods and bootstrapping linguistic resources in languages where no parallel data is available.

3. Constraint-Based Models of Lexical Borrowing

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 patterns from minimal training data. We therefore propose a model whose features are motivated by linguistic knowledge—rather than overparameterized with numerous weakly correlated features which are more practical when large amounts of training data is available. The features in our scoring model are inspired by Optimality Theory (OT; §3.1), 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 illustrated in Figure 2, our model is conceptually divided into three main parts: (1) a mapping of orthographic word forms in two languages into a common phonetic space; (2) generation of loanword pronunciation candidates from a donor word; and (3) ranking of generated loanword candidates, based on linguistic constraints of the donor and recipient languages. In our proposed system, parts (1) and (2) are rule-based; whereas (3) is learned. Each component of the model is discussed in detail in the rest of this section.

The model is implemented as a cascade of finite-state transducers. 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.

Our model performs 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

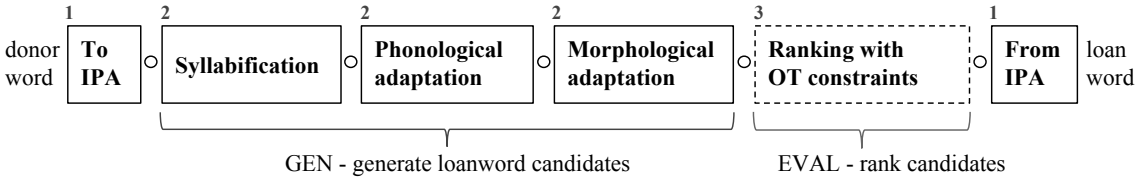

Figure 2: Our morpho-phonological borrowing model conceptually has three main parts: (1) conversion of orthographic word forms to pronunciations in IPA format; (2) generation of loanword pronunciation candidates; (3) ranking of generated candidates using Optimality-Theoretic constraints. Part (1) and (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.

given recipient forms. In a probabilistic cascade, Bayes' rule could be used to reverse the direction and infer underlying donor forms given a loanword. However, we instead opt to train the model discriminatively to find the most likely underlying form, 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 the "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.

3.1 OT: Constraint-Based Evaluation

Borrowing relations may be the result of quite complex transformations on the surface. 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).

OT posits that surface phonetic words of a language emerge from *underlying phonological forms* according to a two-stage process: first, various candidates for the surface form are enumerated for consideration (the 'generation' or GEN phase); then, these candidates are weighed against one another to see which most closely conforms to—or equivalently, least egregiously violates—the phonological preferences of the language. If the preferences are correctly characterized, then the actual surface form should be selected as the most optimal realization of the underlying form. Such preferences are expressed as violable constraints ('violable' because in many cases there may be no candidate that satisfies all of them).

There are two types of OT constraints: *markedness* and *faithfulness* constraints. Markedness constraints (McCarthy & Prince, 1995) describe unnatural (dispreferred) patterns in the language. Faithfulness constraints (Prince & Smolensky, 2008) reward correspondences

/ɛg/		DEP-IO	MAX-IO	ONSET	NO-CODA
a. ☞	ɛg			∗	∗
b.	ɛgə	∗!		∗	
c.	ɛ		∗!	∗	
d.	ʔɛg	∗!			∗

Table 1: A constraint tableau. DEP-IO » MAX-IO » ONSET » NO-CODA are ranked OT constraints according to the phonological system of English. /ɛg/ is the underlying phonological form, and (a), (b) (c), and (d) are the output candidates under consideration. The actual surface form is (a), as it incurs fewer violations than other candidates.

[ʃarr]		*COMPLEX	NO-CODA	MAX-IO	DEP-IO
a.	ʃarr	∗!	∗		
b.	ʃar.ri		∗!		∗
c. ☞	ʃa.ri			∗	∗
d.	ʃa.rri	∗!			∗

Table 2: An example OT analysis adapted to account for borrowing. OT constraints are ranked according to the phonological system of the recipient language (here, Swahili). The donor (Arabic) word شرّ (*$r~*) 'evil' is considered as the underlying form. The winning surface form is the Swahili loanword *shari* 'evil'.

between the underlying form and the surface candidates.[4] As originally proposed, OT holds that the set of constraints is universal, but their ranking is language-specific.

In OT, then, the "grammar" is the set of universal constraints and their language-specific ranking, and a "derivation" for a surface form consists of its underlying form, surface candidates, and constraint violations by those candidates (under which the surface form is correctly chosen). An example of OT analysis is shown in Table 1. Originally OT has been introduced as a generative model of phonology, but later extensions to syntax and semantics turned it into a general theory of grammar.

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. We show an example in Table 2. There has been substantial prior work in linguistics on borrowing in the OT paradigm (Yip, 1993; Davidson & Noyer, 1997; Jacobs & Gussenhoven, 2000; Kang, 2003; Broselow, 2004; Adler, 2006; Rose & Demuth, 2006; Kenstowicz & Suchato, 2006; Kenstowicz, 2007; Mwita, 2009), but none of it has led to computational realizations.

4. 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 how well an output candidate is appropriate to the input), while markedness constraints are analogical to the language model (requiring well-formedness of the output candidate). Without faithfulness constraints, the optimal surface form could differ arbitrarily from the underlying form.

OT assumes an ordinal constraint ranking and strict dominance rather than constraint "weighting". In that, our OT-inspired model departs from OT's standard evaluation assumptions: following Goldwater and Johnson (2003), we use a linear scoring scheme.

3.2 Case Study: Arabic–Swahili Borrowing

In this section, we use the Arabic–Swahili[5] language-pair to describe the prototypical linguistic adaptation processes that words undergo when borrowed. Then, we describe how we model these processes in more general terms.

The Swahili lexicon has been influenced by Arabic due to a prolonged period of language contact due to Indian Ocean trading (800 CE–1920), as well as the influence of Islam (Rothman, 2002). According to several independent studies, Arabic loanwords constitute from 18% (Hurskainen, 2004b) 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. We survey these systems briefly since they illustrate how Arabic loanwords have been substantially adapted to conform to Swahili phonotactics. First, Arabic has five syllable patterns:[6] CV, CVV, CVC, CVCC, and CVVC (McCarthy, 1985, pp. 23–28), whereas Swahili (like other Bantu languages) has only open syllables of the form CV or V. 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).

3.3 Arabic–Swahili Borrowing Transducers

We use unweighted transducers for pronunciation, syllabification, and morphological and phonological adaptation and describe these here. An example that illustrates some of the possible string transformations by individual components of the model is shown in Figure 3. 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

5. For simplicity, we subsume Omani Arabic and other historical dialects of Arabic under the label "Arabic"; our data and examples are in Modern Standard Arabic. Similarly, we subsume Swahili, its dialects and protolanguages under "Swahili".
6. C stands for consonant, and V for vowel.

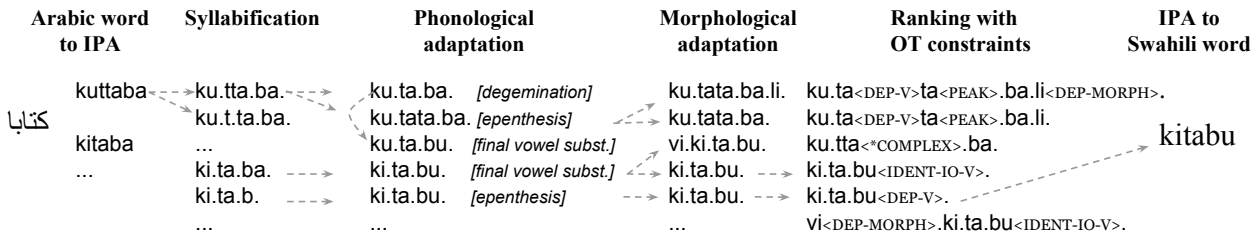

Figure 3: An example of an Arabic word كتابا (*ktAbA*) 'book.sg.indef' transformed by our model into a Swahili loanword *kitabu*.

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.[7] In Arabic, we use the CMU Arabic vowelized pronunciation dictionary containing about 700K types which has an average of four pronunciations per unvowelized input word type (Metze, Hsiao, Jin, Nallasamy, & Schultz, 2010).[8] 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.

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 adaptations.

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 §3.2. 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 /tˤ/, /dˤ/, /sˤ/ 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.ta.tɑ.bɑ/). Sometimes it is possible to predict the final vowel of a word, depending on the word-final

7. www.omniglot.com
8. 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/.

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, Graff, Bouziri, Krouna, & Kulick, 2010); the Swahili affixes are taken from a hand-crafted Swahili morphological analyzer (Littell, Price, & Levin, 2014). For the sake of simplicity in implementation, we strip no more than one Arabic prefix and no more than one suffix per word; and in Swahili – we concatenate at most two Swahili prefixes and at most one suffix.

3.4 Learning Constraint Weights

Due to the computational problems of working with OT (Eisner, 1997, 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 §3.3; (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 with strict dominance (Boersma & Hayes, 2001; Goldwater & Johnson, 2003).

As discussed in §3.1, OT distinguishes markedness constraints which detect dispreferred phonetic patterns in the language, and faithfulness constraints, which ensure correspondences between the underlying form and the surface candidates. The implemented constraints are listed in Tables 3 and 4. 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 §3.3 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 & Mead, 1965), a heuristic derivative-free method that iteratively optimizes, based on an objective function evaluation, the convex hull

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 substitution in manner of pronunciation
IDENT-IO-C-A	no substitution in place of articulation
IDENT-IO-C-S	no substitution 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 substitution in vowel openness
IDENT-IO-V-R	no substitution in vowel roundness
IDENT-IO-V-F	no substitution in vowel frontness
IDENT-IO-V-FIN	no final vowel substitution

Table 3: 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 4: Markedness constraints impose language-specific structural well-formedness of surface realizations.

of $n + 1$ simplex vertices.[9] The objective function used in this work 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.

3.5 Adapting the Model to a New Language

The Arabic–Swahili case study shows that, in principle, a borrowing model can be constructed. But a reasonable question to ask is: how much work is required to build a similar system for a new language pair? We claim that our design permits rapid development in new language pairs. First, 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-trivial problem in some cases, the problem is well studied, and many under-resourced languages (e.g., Swahili), have "phonographic" systems where orthography corresponds to phonology.[10]

To illustrate the ease with which a language pair can be engineered, we applied our borrowing model to the Italian–Maltese and French–Romanian language pairs. Maltese and Romanian, like Swahili, have a large number of borrowed words in their lexicons (Tadmor, 2009). Maltese (a phylogenetically Semitic language) has 35.1%–30.3% loanwords of Romance (Italian/Sicilian) origin (Comrie & Spagnol, 2015). Although French and Romanian are sister languages (both descending from Latin), about 12% of Romanian types are French borrowings that came into the language in the past few centuries (Schulte, 2009). For both language pairs we manually define a set of allowed insertions, deletions, and substitutions of phonemes and morphemes, based on the training sets. A set of Maltese affixes was defined based on the linguistic survey by Fabri, Gasser, Habash, Kiraz, and Wintner (2014). We employ the GlobalPhone pronunciation dictionary for French (Schultz & Schlippe, 2014), converted to IPA, and automatically constructed Italian, Romanian, and Maltese pronunciation dictionaries using the Omniglot grapheme-to-IPA conversion rules for those languages.

4. Models of Lexical Borrowing in Statistical Machine Translation

Before turning to an experimental verification and analysis of the borrowing model, we introduce an external application where the borrowing model will be used as a component—machine translation. We rely on the borrowing model to project translation information

9. The decision to use Nelder–Mead rather than more conventional gradient-based optimization algorithms was motivated purely by practical limitations of the finite-state toolkit we used which made computing derivatives with latent structure impractical from an engineering standpoint.

10. This tendency can be explained by the fact that, in many cases, lower-resource languages have developed orthography relatively recently, rather than having organically evolved written forms that preserve archaic or idiosyncratic spellings that are more distantly related to the current phonology of the language such as we see in, e.g., English.

from a high-resource donor language into a low-resource recipient language, thus mitigating the deleterious effects of out-of-vocabulary (OOV) words.

OOVs are a ubiquitous and difficult problem in MT. When a translation system encounters an OOV—a word that was not observed in the training data, and the trained system thus lacks its translation variants—it usually outputs the word just as it is in the source language, producing erroneous and disfluent translations. All MT systems, even when trained on billion-sentence-size parallel corpora, will encounter OOVs at test time. Often, these are named entities and neologisms. However, the OOV problem is much more acute in morphologically-rich and low-resource scenarios: there, OOVs are primarily not lexicon-peripheral items such as names and specialized/technical terms, but also regular content words. Since borrowed words are a component of the regular lexical content of a language, projecting translations onto the recipient language by identifying borrowed lexical material is a plausible strategy for solving this problem.

Procuring translations for OOVs has been a subject of active research for decades. Translation of named entities is usually generated using transliteration techniques (Al-Onaizan & Knight, 2002; Hermjakob, Knight, & Daumé III, 2008; Habash, 2008). Extracting a translation lexicon for recovering OOV content words and phrases is done by mining bi-lingual and monolingual resources (Rapp, 1995; Callison-Burch, Koehn, & Osborne, 2006; Haghighi, Liang, Berg-Kirkpatrick, & Klein, 2008; Marton, Callison-Burch, & Resnik, 2009; Razmara, Siahbani, Haffari, & Sarkar, 2013; Saluja, Hassan, Toutanova, & Quirk, 2014). In addition, OOV content words can be recovered by exploiting cognates, by transliterating and then "pivoting" via a closely-related resource-richer language, when such a language exists (Hajič, Hric, & Kuboň, 2000; Mann & Yarowsky, 2001; Kondrak, Marcu, & Knight, 2003; De Gispert & Marino, 2006; Habash & Hu, 2009; Durrani, Sajjad, Fraser, & Schmid, 2010; Wang, Nakov, & Ng, 2012; Nakov & Ng, 2012; Dholakia & Sarkar, 2014). Our work is similar in spirit to the latter pivoting approach, but we show how to obtain translations for OOV content words by pivoting via an unrelated, often typologically distant resource-rich language.

Our solution is depicted, at a high level, in Figure 4. Given an OOV word in resource-poor MT, we use our borrowing system to identify list of likely donor words from the donor language. Then, using the MT system in the resource-rich language, we translate the donor words to the same target language as in the resource-poor MT (here, English). Finally, we integrate translation candidates in the resource-poor system.

We now discuss integrating translation candidates acquired via borrowing plus resource-rich translation.

Briefly, phrase-based translation works as follows. A set of candidate translations for an input sentence is created by matching contiguous spans of the input against an inventory of phrasal translations, reordering them into a target-language appropriate order, and choosing the best one according to a model that combines features of the phrases used, reordering patterns, and target language model (Koehn, Och, & Marcu, 2003). A limitation of this approach is that it can only generate input/output phrase pairs that were directly observed in the training corpus. In resource-limited languages, the standard phrasal inventory will generally be incomplete due to limited parallel data. Thus, the decoder's only hope for producing a good output is to find a fluent, meaning-preserving translation using incomplete translation lexicons. "Synthetic phrases" (Tsvetkov, Dyer, Levin, & Bhatia, 2013; Chahuneau,

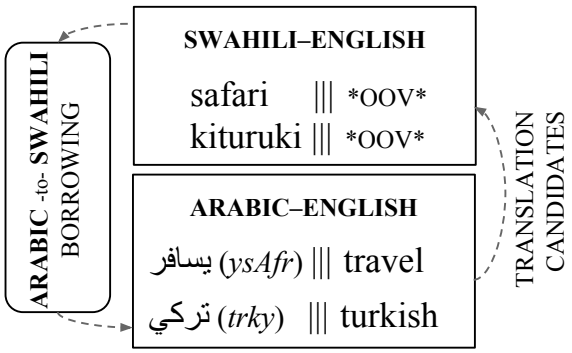

Figure 4: To improve a resource-poor Swahili–English MT system, we extract translation candidates for OOV Swahili words borrowed from Arabic using the Swahili-to-Arabic borrowing system and Arabic–English resource-rich MT.

Schlinger, Smith, & Dyer, 2013; Schlinger, Chahuneau, & Dyer, 2013; Ammar, Chahuneau, Denkowski, Hanneman, Ling, Matthews, Murray, Segall, Tsvetkov, Lavie, & Dyer, 2013; Tsvetkov, Metze, & Dyer, 2014) is a strategy of integrating translated phrases directly in the MT *translation model*, rather than via pre- or post-processing MT inputs and outputs. Synthetic phrases are phrasal translations that are not directly extractable from the training data, generated by auxiliary translation and postediting processes (for example, extracted from a borrowing model). An important advantage of synthetic phrases is that the process often benefits from phrase synthesizers that have high recall (relative to precision) since the global translation model will still have the final say on whether a synthesized phrase will be used.

For each OOV, the borrowing system produces the n-best list of plausible donors; for each donor we then extract the k-best list of its translations.[11] Then, we pair the OOV with the resulting $n \times k$ translation candidates. The translation candidates are noisy: some of the generated donors may be erroneous, the errors are then propagated in translation.[12] To allow the low-resource translation system to leverage good translations that are missing in the default phrase inventory, while being able to learn how trustworthy they are, we integrate the borrowing-model acquired translation candidates as synthetic phrases.

To let the translation model learn whether to trust these phrases, the translation options obtained from the borrowing model are augmented with an indicator feature indicating that the phrase was generated externally (i.e., rather than being extracted from the parallel data). Additional features assess properties of the donor–loan words' relation; their goal is to provide an indication of plausibility of the pair (to mark possible errors in the outputs of the borrowing system). We employ two types of features: phonetic and semantic. Since borrowing is primarily a phonological phenomenon, phonetic features will provide an indication of how

11. We set n and k to 5, we did not experiment with other values.
12. We give as input into the borrowing system all OOV words, although, clearly, not all OOVs are loanwords, and not all loanword OOVs are borrowed from the donor language. However, an important property of the borrowing model is that its operations are not general, but specific to the language-pair and reduced only to a small set of plausible changes that the donor word can undergo in the process of assimilation in the recipient language. Thus, the borrowing system only *minimally* overgenerates the set of output candidates given an input. If the borrowing system encounters an input word that was not borrowed from the target donor language, it usually (but not always) produces an empty output.

typical (or atypical) pronunciation of the word in a language; loanwords are expected to be less typical than core vocabulary words. The goal of semantic features is to measure semantic similarity between donor and loan words: erroneous candidates and borrowed words that changed meaning over time are expected to have different meaning from the OOV.

Phonetic features. To compute phonetic features we first train a (5-gram) language model (LM) of IPA pronunciations of the donor/recipient language vocabulary (p_ϕ). Then, we re-score pronunciations of the donor and loanword candidates using the LMs. We hypothesize that in donor–loanword pairs the donor phoneLM score is higher but the loanword score is lower (i.e., the loanword phonology is atypical in the recipient language). We capture this intuition in three features: $f_1 = p_\phi(\textit{donor})$, $f_2 = p_\phi(\textit{loanword})$, and the harmonic mean between the two scores $f_3 = \frac{2f_1f_2}{f_1+f_2}$.

Semantic features. We compute a semantic similarity feature between the candidate donor and the OOV loanword as follows. We first train, using large monolingual corpora, 100-dimensional word vector representations for donor and recipient language vocabularies.[13] Then, we employ canonical correlation analysis (CCA) with small donor–loanword dictionaries (training sets in the borrowing models) to project the word embeddings into 50-dimensional vectors with maximized correlation between their dimensions. The semantic feature annotating the synthetic translation candidates is cosine distance between the resulting donor and loanword vectors. We use the `word2vec` Skip-gram model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to train monolingual vectors,[14] and the CCA-based tool (Faruqui & Dyer, 2014) for projecting word vectors.[15]

5. Experiments

We now turn to the problem of empirically validating the model we have proposed. Our evaluation consists of two parts. First, we perform an intrinsic assessment of the model's ability to learn borrowing correspondences and compare these to similar approaches that use less linguistic knowledge but which have been used to solve similar string mapping problems. Second, we show the effect of borrowing-augmented translations in translation systems, exploring the effects of the features proposed above.

5.1 Intrinsic Evaluation of Models of Lexical Borrowing

Our experimental setup is defined as follows. The input to the borrowing model is a loanword candidate in Swahili/Maltese/Romanian, the outputs are plausible donor words in the Arabic/Italian/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.

13. We assume that while parallel data is limited in the recipient language, monolingual data is available.
14. `code.google.com/p/word2vec`
15. `github.com/mfaruqui/eacl14-cca`

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.[16] 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, Chahuneau, & Smith, 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, Rambow, & Roth, 2009) for Arabic morphological expansion.

From the resulting dataset of 490 extracted Arabic–Swahili borrowing examples,[17] we set aside randomly sampled 73 examples (15%) for evaluation,[18] and use the remaining 417 examples for model parameter optimization. For Italian–Maltese language pair, we use the same technique and extract 425 training and 75 (15%) randomly sampled test examples. For French–Romanian language pair, we use an existing small annotated set of borrowing examples,[19] with 282 training and 50 (15%) randomly sampled test examples.

We use `pyfst`—a Python interface to OpenFst (Allauzen, Riley, Schalkwyk, Skut, & Mohri, 2007)—for the borrowing model implementation.[20]

Baselines. We compare our model to several baselines. In the Levenshtein distance baselines we chose the closest word (either surface or pronunciation-based). In the cognates baselines, we evaluate a variant of the Levenshtein distance tuned to identify cognates (Mann & Yarowsky, 2001; Kondrak & Sherif, 2006); this method was identified by Kondrak and Sherif (2006) among the top three cognate identification methods. In the transliteration baselines we generate plausible transliterations of the input Swahili (or Romanian) words in the donor lexicon using the model of Ammar, Dyer, and Smith (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 conjoined with a moving window of source characters. In the OT-uniform baselines, we evaluate the accuracy of the borrowing model with uniform weights, thus the shortest path in the loanwords transducer will be forms that violate 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

16. 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.
17. In each training/test example one Swahili word corresponds to all extracted Arabic donor words.
18. 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'.
19. `http://wold.clld.org/vocabulary/8`
20. `https://github.com/vchahun/pyfst`

	AR–SW	IT–MT	FR–RO
Reachability	87.7%	92.7	82.0%
Ambiguity	857	11	12

Table 5: 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	IT–MT	FR–RO
Orthographic baselines	Levenshtein-orthographic	8.9	61.5	38.0
	Transliteration	16.4	61.3	36.0
Phonetic baselines	Levenshtein-pronunciation	19.8	64.4	26.3
	Cognates	19.7	63.7	30.7
OT	OT-uniform constraint weights	29.3	65.6	58.5
	OT-learned constraint weights	**48.4**	**83.3**	**75.6**

Table 6: The evaluation of the borrowing model accuracy. We compare the following setups: orthographic (surface) and phonetic (based on pronunciation lexicon) Levenshtein distance, a cognate identification model that uses heuristic Levenshtein distance with lower penalty on vowel updates and similar letter/phone substitutions, a CRF transliteration model, and our model with uniform and learned OT constraint weights assignment.

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; however, inference may be expensive and the discriminative component will have a greater burden. 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.[21]

Results. The reachability and ambiguity of the borrowing model are listed in Table 5. Briefly, the model obtains high reachability, while significantly reducing the average number of possible outputs per input: in Arabic from 787K to 857 words, in Maltese from 129K to 11, 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 6 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

21. Our measure of ambiguity is equivalent to perplexity assuming a uniform distribution over output forms.

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.

EN	AR orth.	AR pron.	SW syl.	Violated 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 7: Examples of inferred syllabification and corresponding constraint violations produced by our borrowing model.

Qualitative evaluation. The constraint ranking learned by the borrowing model (constraints are listed in Tables 3, 4) is in line with prior linguistic analysis. In Swahili NO-CODA dominates all other markedness constraints. 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. Morphologically-motivated constraints also comply with tendencies discussed in linguistic literature: donor words may remain unmodified and are treated as a stem, and then are reinfected according to the recipient morphology, thus DEP-IO-MORPH can be dominated more easily than MAX-IO-MORPH. 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 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[22] along with the constraint violations produced by the borrowing model are depicted in Table 7.

5.2 Extrinsic Evaluation of Pivoting via Borrowing in MT

We now turn to an extrinsic evaluation, looking at two low-resource translation tasks: Swahili–English translation (resource-rich donor language: Arabic), Maltese–English translation (resource-rich donor language: Italian), and Romanian–English translation (resource-rich donor language: French). We begin by reviewing the datasets used, and then discuss two oracle experiments that attempt to quantify how much value could we obtain from a perfect borrowing model (since not all mistakes made by MT systems involve borrowed words). Armed with this understanding, we then explore how much improvement can be obtained using our system.

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

Datasets and software. The Swahili–English parallel corpus was crawled from the Global Voices project website[23]. For the Maltese–English language pair, we sample a parallel corpus of the same size from the EUbookshop corpus from the OPUS collection (Tiedemann, 2012). Similarly, to simulate resource-poor scenario for the Romanian–English language pair, we sample a corpus from the transcribed TED talks (Cettolo, Girardi, & Federico, 2012). To evaluate translation improvement on corpora of different sizes we conduct experiments with sub-sampled 4K, 8K, and 14K parallel sentences from the training corpora (the smaller the training corpus, the more OOVs it has). Corpora sizes along with statistics of source-side OOV tokens and types are given in Table 8. Statistics of the held-out dev and test sets used in all translation experiments are given in Table 9.

Arabic–English pivot translation system was trained on a parallel corpus of about 5.4M sentences available from the Linguistic Data Consortium (LDC), and optimized on the standard NIST MTEval dataset for the year 2005 (MT05). Italian–English system was trained on 11M sentences from the OPUS corpus. French–English pivot system was trained on about 400K sentences from the transcribed TED talks, and optimized on the dev talks from the Romanian–English system; test talks from the Romanian–English system were removed from the French–English training corpus.

In all the MT experiments, we use the `cdec`[24] toolkit (Dyer, Lopez, Ganitkevitch, Weese, Ture, Blunsom, Setiawan, Eidelman, & Resnik, 2010), and optimize parameters with MERT (Och, 2003). English 4-gram language models with Kneser-Ney smoothing (Kneser & Ney, 1995) were trained using KenLM (Heafield, 2011) on the target side of the parallel training corpora and on the Gigaword corpus (Parker, Graff, Kong, Chen, & Maeda, 2009). Results are reported using case-insensitive BLEU with a single reference (Papineni, Roukos, Ward, & Zhu, 2002). To verify that our improvements are consistent and are not just an effect of optimizer instability, we train three systems for each MT setup; reported BLEU scores are averaged over systems.

Upper bounds. The goal of our experiments is not only to evaluate the contribution of the OOV dictionaries that we extract when pivoting via borrowing, but also to understand the potential contribution of exploiting borrowing.What is the overall improvement that would be achieved if we could correctly translate all OOVs that were borrowed from another language? What is the overall improvement that can be achieved if we correctly translate all OOVs? We answer this question by defining "upper bound" experiments. In the upper bound experiment we word-align all available parallel corpora, including dev and test sets, and extract from the alignments oracle translations of OOV words. Then, we append the extracted OOV dictionaries to the training corpora and re-train SMT setups without OOVs. Translation scores of the resulting system provide an upper bound of an improvement from correctly translating all OOVs. When we append oracle translations of the subset of OOV dictionaries, in particular translations of all OOVs for which the output of the borrowing system is not empty, we obtain an upper bound that can be achieved using our method (if the borrowing system provided perfect outputs relative to the reference translations). Understanding the upper bounds is relevant not only for our experiments, but for any experiments that involve augmenting translation dictionaries; however, we are not aware

23. `sw.globalvoicesonline.org`
24. `www.cdec-decoder.org`

		4K	8K	14K
SW–EN	Tokens	84,764	170,493	300,648
	Types	14,554	23,134	33,288
	OOV tokens	4,465 (12.7%)	3,509 (10.0%)	2,965 (8.4%)
	OOV types	3,610 (50.3%)	2,950 (41.1%)	2,523 (35.1%)
MT–EN	Tokens	104,181	206,781	358,373
	Types	14,605	22,407	31,176
	OOV tokens	4,735(8.7%)	3,497 (6.4%)	2,840 (5.2%)
	OOV types	4,171 (44.0%)	3,236 (34.2%)	2,673 (28.2%)
RO–EN	Tokens	35,978	71,584	121,718
	Types	7,210	11,144	15,112
	OOV tokens	3,268 (16.6%)	2,585 (13.1%)	2,177 (11.1%)
	OOV types	2,382 (55.0%)	1,922 (44.4%)	1,649 (38.1%)

Table 8: Statistics of the Swahili–English, Maltese–English, and Romanian–English corpora and source-side OOV rates for 4K, 8K, 14K parallel training sentences.

	SW–EN		MT–EN		RO–EN	
	dev	test	dev	test	dev	test
Sentences	1,552	1,732	2,000	2,000	2,687	2,265
Tokens	33,446	35,057	54,628	54,272	24,754	19,659
Types	7,008	7,180	9,508	9,471	5,141	4,328

Table 9: Dev and test corpora sizes.

of prior work providing similar analysis of upper bounds, and we recommend this as a calibrating procedure for future work on OOV mitigation strategies.

Borrowing-augmented setups. As described in §4, we integrate translations of OOV loanwords in the translation model using the synthetic phrase paradigm. Due to data sparsity, we conjecture that non-OOVs that occur only few times in the training corpus can also lack appropriate translation candidates, i.e., these are target-language OOVs. We therefore plug into the borrowing system OOVs and non-OOV words that occur less than 3 times in the training corpus. We list in Table 10 sizes of resulting borrowed lexicons that we integrate in translation tables.

	4K	8K	14K
Loan OOVs in SW–EN	5,050	4,219	3,577
Loan OOVs in MT–EN	10,138	6,456	4,883
Loan OOVs in RO–EN	347	271	216

Table 10: Sizes of dictionaries extracted using pivoting via borrowing and integrated in translation models.

Transliteration-augmented setups. In addition to the standard baselines, we evaluate transliteration baselines, where we replace the borrowing model by the baselines described in §5.1. As in the borrowing system, transliteration outputs are filtered to contain only target language lexicons. We list in Table 11 sizes of obtained translated lexicons.

	4K	8K	14K
Transliteration OOVs in SW–EN	49	32	22
Transliteration OOVs in MT–EN	26,734	19,049	15,008
Transliteration OOVs in RO–EN	906	714	578

Table 11: Sizes of translated lexicons extracted using pivoting via transliteration and integrated in translation models.

Results. Translation results are shown in Table 12. We evaluate separately the contribution of the integrated OOV translations, and the same translations annotated with phonetic and semantic features. We also provide upper bound scores for integrated loanword dictionaries as well as for recovering all OOVs.

		4K	8K	14K
SW–EN	Baseline	13.2	15.1	17.1
	+ Transliteration OOVs	13.4	15.3	17.2
	+ Loan OOVs	14.3	15.7	18.2
	+ Features	**14.8**	**16.4**	**18.4**
	Upper bound loan	18.9	19.1	20.7
	Upper bound all OOVs	19.2	20.4	21.1
MT–EN	Baseline	26.4	31.4	35.2
	+ Transliteration OOVs	26.5	30.8	34.9
	+ Loan OOVs	**27.2**	31.7	**35.3**
	+ Features	26.9	**31.9**	34.5
	Upper bound loan	28.5	32.2	35.7
	Upper bound all OOVs	31.6	35.6	38.0
RO–EN	Baseline	15.8	18.5	20.7
	+ Transliteration OOVs	15.8	**18.7**	**20.8**
	+ Loan OOVs	**16.0**	**18.7**	20.7
	+ Features	**16.0**	18.6	20.6
	Upper bound loan	16.6	19.4	20.9
	Upper bound all OOVs	28.0	28.8	30.4

Table 12: Swahili–English, Maltese–English, and Romanian–English MT experiments.

Swahili–English MT performance is improved by up to +1.6 BLEU when we augment it with translated OOV loanwords leveraged from the Arabic–Swahili borrowing and then Arabic–English MT. The contribution of the borrowing dictionaries is +0.6–1.1 BLEU, and phonetic and semantic features contribute additional half BLEU. More importantly, upper bound results show that the system can be improved more substantially with better

dictionaries of OOV loanwords. This result confirms that OOV borrowed words is an important type of OOVs, and with proper modeling it has the potential to improve translation by a large margin. Maltese–English system is also improved substantially, by up to +0.8 BLEU, but the contribution of additional features is less pronounced. Romanian–English systems obtain only small but significant improvement for 4K and 8K, $p < .01$ (Clark, Dyer, Lavie, & Smith, 2011). However, this is expected as the rate of borrowing from French into Romanian is smaller, and, as the result, the integrated loanword dictionaries are small. Transliteration baseline, conversely, is more effective in Romanian–French language pair, as the languages are related typologically, and have common cognates in addition to loanwords. Still, even with these dictionaries the translations with pivoting via borrowing/transliteration improve, and even almost approach the upper bounds results.

Error analysis. Our augmented MT systems combine three main components: the translation system itself, a borrowing system, and a pivot translation system. At each step of the application errors may occur that lead to erroneous translation. To identify main sources of errors in the Swahili–English end-to-end system, we conducted a manual analysis of errors in translations of OOV types produced by the Swahili–English 4K translation systems. As a gold standard corpus we use the Helsinki Corpus of Swahili[25] (Hurskainen, 2004a, HCS). HCS is a morphologically, syntactically, and semantically annotated corpus of about 580K sentences (12.7M tokens). In the corpus 52,351 surface forms (1.5M tokens) are marked as Arabic loanwords. Out of the 3,610 OOV types in the Swahili–English 4K translation systems, 481 word types are annotated in the HCS. We manually annotated these 481 words and identified 353 errors; the remaining 128 words were translated correctly in the end-to-end system. Our analysis reveals the error sources detailed below. In Table 13 we summarize the statistics of the error sources.

Error source	#	%
Reachability of the borrowing system	113	32.0
Loanword production errors	191	54.1
Arabic–English translation errors	20	5.7
Swahili–English translation errors	29	8.2

Table 13: Sources of errors.

1. **Reachability of the borrowing system.**
   Only 368 out of 481 input words produced loanword candidates. The main reason for the unreachable paths is complex morphology of Swahili OOVs, not taken into account by our borrowing system. For example, *atakayehusika* 'who will be involved', the lemma is *husika* 'involve'.

2. **Loanword production errors.**
   About half of errors are due to incorrect outputs of the borrowing system. This is in line with the Arabic–Swahili borrowing system accuracy reported in Table 6. For example, all morphological variants of the lemma *wahi* 'never' (*hayajawahi*, *halijawahi*,

25. www.aakkl.helsinki.fi/cameel/corpus/intro.htm

hazijawahi), incorrectly produced an Arabic donor word جاوه (*jAwh*) 'java'. Additional examples include all variants of the lemma *saidia* 'help' (*isaidie*, *kimewasaidia*) produced Arabic donor candidates that are variants of the proper name *Saidia*.

3. **Arabic–English translation errors.**
   As the most frequent source of errors in the Arabic–English MT system, we have identified OOV Arabic words. For example, although for the Swahili loanword *awashukuru* 'thank you' the borrowing system correctly produced a plausible donor word وشكور (*w$kwr*) 'and thank you' (rarely used), the only translation variant produced by the Arabic–English MT was *kochkor*.

4. **Swahili–English translation errors.**
   In some cases, although the borrowing system produced a correct donor candidate, and the Arabic–English translation was also correct, translation variants were different from the reference translations in the Swahili–English MT system. For example, the word *alihuzunika* 'he grieved' correctly produced an Arabic donor الحزن (*AlHzn*) 'grief'. Translation variants produced by the Arabic–English MT are *sadness, grief, saddened, sorrow, sad, mourning, grieved, saddening, mourn, distressed*, whereas the expected translation in the Swahili–English reference translations is *disappointed*. Another source of errors that occurred despite correct outputs of borrowing and translation systems is historical meaning change of words. An interesting example of such semantic shift is the word *sakafu* 'floor', that was borrowed from the Arabic word سقف (*sqf*) 'ceiling'.

Complex morphology of both Swahili and Arabic is the most frequent source of errors at all steps of the application. Concatenation of several prefixes in Swahili affects the reachability of the borrowing system. Some Swahili prefixes flip the meaning of words, e.g. *kutoadhibiwa* 'impunity', produces the lemma *adhibiwa* 'punishment', and consequently translations *torture, torturing, tortured*. Finally, derivational processes in both languages are not handled by our system, e.g., a verb *aliyorithi* 'he inherited', produces an Arabic noun الوارثة (*AlwArvp*) 'the heiress', and its English translations *heiress*. Jointly reasoning about morphological processes in the donor and recipient languages suggests a possible avenue for remedying these issues.

6. Additional 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."

7. 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. Our translation results substantially improve over the baseline and confirm that OOV loanwords are important and merit further investigation.

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 additional 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.

References

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

Ahn, S.-C., & Iverson, G. K. (2004). Dimensions in Korean laryngeal phonology. *Journal of East Asian Linguistics*, *13*(4), 345–379.

Al-Onaizan, Y., & Knight, K. (2002). Machine transliteration of names in Arabic text. In *Proc. the ACL workshop on Computational Approaches to Semitic Languages*, pp. 1–13. Association for Computational Linguistics.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pp. 11–23. Springer.

Ammar, W., Chahuneau, V., Denkowski, M., Hanneman, G., Ling, W., Matthews, A., Murray, K., Segall, N., Tsvetkov, Y., Lavie, A., & Dyer, C. (2013). The cmu machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proc. WMT*.

Ammar, W., Dyer, C., & Smith, N. A. (2012). Transliteration by sequence labeling with lattice encodings and reranking. In *Proc. NEWS workshop at ACL*.

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

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, *32*(1), 45–86.

Broselow, E. (2004). Language contact phonology: richness of the stimulus, poverty of the base. In *Proc. NELS*, Vol. 34, pp. 1–22.

Burkett, D., & Klein, D. (2008). Two languages are better than one (for syntactic parsing). In *Proc. EMNLP*, pp. 877–886.

Calabrese, A., & Wetzels, W. L. (2009). *Loan phonology*, Vol. 307. John Benjamins Publishing.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proc. ACL*.

Cettolo, M., Girardi, C., & Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proc. EAMT*, pp. 261–268.

Chahuneau, V., Schlinger, E., Smith, N. A., & Dyer, C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proc. EMNLP*, pp. 1677–1687.

Clark, J. H., Dyer, C., Lavie, A., & Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. ACL*, pp. 176–181.

Comrie, B., & Spagnol, M. (2015). Maltese loanword typology. Submitted.

Das, D., & Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. ACL*, pp. 600–609. Association for Computational Linguistics.

Daumé III, H. (2009). Non-parametric Bayesian areal linguistics. In *Proc. NAACL*, pp. 593–601. Association for Computational Linguistics.

Davidson, L., & Noyer, R. (1997). Loan phonology in Huave: nativization and the ranking of faithfulness constraints. In *Proc. WCCFL*, Vol. 15, pp. 65–79.

De Gispert, A., & Marino, J. B. (2006). Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proc. LREC*, pp. 65–68.

Dholakia, R., & Sarkar, A. (2014). Pivot-based triangulation for low-resource languages. In *Proc. AMTA*.

Diab, M., & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proc. ACL*.

Durrani, N., Sajjad, H., Fraser, A., & Schmid, H. (2010). Hindi-to-Urdu machine translation through transliteration. In *Pro. ACL*, pp. 465–474.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. NAACL*.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., & Resnik, P. (2010). `cdec`: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. ACL*.

Eisner, J. (1997). Efficient generation in primitive Optimality Theory. In *Proc. EACL*, pp. 313–320.

Eisner, J. (2002). Comprehension and compilation in Optimality Theory. In *Proc. ACL*, pp. 56–63.

Ellison, T. M. (1994). Phonological derivation in Optimality Theory. In *Proc. CICLing*, pp. 1007–1013.

Fabri, R., Gasser, M., Habash, N., Kiraz, G., & Wintner, S. (2014). Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In *Natural Language Processing of Semitic Languages*, pp. 3–41. Springer.

Faruqui, M., & Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proc. EACL*.

Ganchev, K., Gillenwater, J., & Taskar, B. (2009). Dependency grammar induction via bitext projection constraints. In *Proc. ACL*, pp. 369–377. Association for Computational Linguistics.

Garley, M., & Hockenmaier, J. (2012). Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proc. ACL*, pp. 135–139.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proc. the Stockholm workshop on variation within Optimality Theory*, pp. 111–120.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proc. ACL*, pp. 57–60.

Habash, N., & Hu, J. (2009). Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proc. WMT*, pp. 173–181.

Habash, N., Rambow, O., & Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proc. MEDAR*, pp. 102–109.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proc. ACL*, pp. 771–779.

Hajič, J., Hric, J., & Kuboň, V. (2000). Machine translation of very close languages. In *Proc. ANLP*, pp. 7–12.

Haspelmath, M. (2009). Lexical borrowing: concepts and issues. *Loanwords in the World's Languages: a comparative handbook*, 35–54.

Haspelmath, M., & Tadmor, U. (Eds.). (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 210–231.

Hayes, B., Tesar, B., & Zuraw, K. (2013). OTSoft 2.3.2..

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proc. WMT*.

Hermjakob, U., Knight, K., & Daumé III, H. (2008). Name translation in statistical machine translation-learning when to transliterate. In *Proc. ACL*, pp. 389–397.

Hock, H. H., & Joseph, B. D. (2009). *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*, Vol. 218. Walter de Gruyter.

Holden, K. (1976). Assimilation rates of borrowings and phonological productivity. *Language*, 131–147.

Hurskainen, A. (2004a). HCS 2004–Helsinki corpus of Swahili. Tech. rep., Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

Hurskainen, A. (2004b). Loan words in Swahili. In Bromber, K., & Smieja, B. (Eds.), *Globalisation and African Languages*, pp. 199–218. Walter de Gruyter.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, *11*(3).

Jacobs, H., & Gussenhoven, C. (2000). Loan phonology: perception, salience, the lexicon and OT. *Optimality Theory: Phonology, syntax, and acquisition*, 193–209.

Johnson, F. (1939). *Standard Swahili-English dictionary*. Oxford University Press.

Kager, R. (1999). *Optimality Theory*. Cambridge University Press.

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

Kang, Y. (2011). Loanword phonology. In van Oostendorp, M., Ewen, C., Hume, E., & Rice, K. (Eds.), *Companion to Phonology*. Wiley–Blackwell.

Kawahara, S. (2008). Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics*, *17*(4), 317–330.

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

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

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, Vol. 1, pp. 181–184. IEEE.

Knight, K., & Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, *24*(4), 599–612.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proc. NAACL-HLT*, pp. 48–54.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In *Proc. NAACL*, pp. 1–8. Association for Computational Linguistics.

Kondrak, G., Marcu, D., & Knight, K. (2003). Cognates can improve statistical translation models. In *Proc. HLT-NAACL*, pp. 46–48. Association for Computational Linguistics.

Kondrak, G., & Sherif, T. (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proc. the Workshop on Linguistic Distances*, pp. 43–50. Association for Computational Linguistics.

Kozhevnikov, M., & Titov, I. (2013). Cross-lingual transfer of semantic role labeling models. In *Proc. ACL*, pp. 1190–1200.

Kuhn, J. (2004). Experiments in parallel-text based grammar induction. In *Proc. ACL*, p. 470.

Li, S., Graça, J. V., & Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proc. EMNLP*, pp. 1389–1398.

List, J.-M., & Moran, S. (2013). An open source toolkit for quantitative historical linguistics. In *Proc. ACL (System Demonstrations)*, pp. 13–18.

Littell, P., Price, K., & Levin, L. (2014). Morphological parsing of Swahili using crowdsourced lexical resources. In *Proc. LREC*.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., & Kulick, S. (2010). LDC Standard Arabic morphological analyzer (SAMA) v. 3.1..

Mann, G. S., & Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In *Proc. HLT-NAACL*, pp. 1–8.

Marton, Y., Callison-Burch, C., & Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proc. EMNLP*, pp. 381–390.

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

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

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

Metze, F., Hsiao, R., Jin, Q., Nallasamy, U., & Schultz, T. (2010). The 2010 CMU GALE speech-to-text system. In *Proc. INTERSPEECH*, pp. 1501–1504.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pp. 3111–3119.

Moravcsik, E. (1978). Language contact. *Universals of human language*, *1*, 93–122.

Mwita, L. C. (2009). The adaptation of Swahili loanwords from Arabic: A constraint-based analysis. *Journal of Pan African Studies*.

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press Oxford.

Nakov, P., & Ng, H. T. (2012). Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 179–222.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer journal*, *7*(4), 308–313.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. ACL*, pp. 160–167.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, *36*(1), 307–340.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pp. 311–318.

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2009). English Gigaword fourth edition..

Polomé, E. C. (1967). *Swahili Language Handbook*. ERIC.

Prince, A., & Smolensky, P. (2008). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Pro. ACL*, pp. 320–322.

Razmara, M., Siahbani, M., Haffari, R., & Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proc. ACL*, pp. 1105–1115.

Repetti, L. (2006). The emergence of marked structures in the integration of loans in Italian. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, *274*, 209.

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

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

Saluja, A., Hassan, H., Toutanova, K., & Quirk, C. (2014). Graph-based semi-supervised learning of translation models from monolingual data. In *Proc. ACL*, pp. 676–686.

Sankoff, G. (2002). Linguistic outcomes of language contact. In Chambers, J., Trudgill, P., & Schilling-Estes, N. (Eds.), *Handbook of Sociolinguistics*, pp. 638–668. Blackwell.

Schadeberg, T. C. (2009). Loanwords in Swahili. In Haspelmath, M., & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook*, pp. 76–102. Max Planck Institute for Evolutionary Anthropology.

Schlinger, E., Chahuneau, V., & Dyer, C. (2013). `morphogen`: Translation into morphologically rich languages with synthetic phrases. *The Prague Bulletin of Mathematical Linguistics*, *100*, 51–62.

Schulte, K. (2009). Loanwords in Romanian. In Haspelmath, M., & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook*, pp. 230–259. Max Planck Institute for Evolutionary Anthropology.

Schultz, T., & Schlippe, T. (2014). GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proc. LREC*.

Smith, D. A., & Smith, N. A. (2004). Bilingual parsing with factored estimation: Using english to parse korean.. In *Proc. EMNLP*, pp. 49–56.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., & Lopez, A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proc. ACL*, pp. 1374–1383.

Snyder, B., Naseem, T., & Barzilay, R. (2009). Unsupervised multilingual grammar induction. In *Proc. ACL/AFNLP*, pp. 73–81.

Täckström, O., Das, D., Petrov, S., McDonald, R., & Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging.. *Transactions of the Association for Computational Linguistics*, *1*, 1–12.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In Haspelmath, M., & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook*, pp. 55–75. Max Planck Institute for Evolutionary Anthropology.

Thomason, S. G., & Kaufman, T. (2001). *Language contact*. Edinburgh University Press Edinburgh.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proc. of LREC*, pp. 2214–2218.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *Proc. COLING*.

Tsvetkov, Y., Ammar, W., & Dyer, C. (2015). Constraint-based models of lexical borrowing. In *Proc. NAACL*, pp. 598–608.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proc. ACL*, pp. 248–258.

Tsvetkov, Y., & Dyer, C. (2015). Lexicon stratification for translating out-of-vocabulary words. In *Proc. ACL*.

Tsvetkov, Y., Dyer, C., Levin, L., & Bhatia, A. (2013). Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. WMT*.

Tsvetkov, Y., Metze, F., & Dyer, C. (2014). Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In *Proc. EACL*, pp. 616–625.

Van Coetsem, F. (1988). *Loan phonology and the two transfer types in language contact*. Walter de Gruyter.

Wang, P., Nakov, P., & Ng, H. T. (2012). Source language adaptation for resource-poor machine translation. In *Proc. EMNLP*, pp. 286–296.

Weinreich, U. (1979). *Languages in contact: Findings and problems*. Walter de Gruyter.

Whitney, W. D. (1881). On mixture in language. *Transactions of the American Philological Association (1870)*, 5–26.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, *23*(3), 377–403.

Xi, C., & Hwa, R. (2005). A backoff model for bootstrapping resources for non-English languages. In *Proc. EMNLP*, pp. 851–858.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. HLT*, pp. 1–8.

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

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