**Loanword Phonology of French Words in Vietnamese Modeled with a Finite State Toolkit**Giang H. Le

University of Illinois at Urbana Champaign

LING 542

Professor Chilin Shih – Department of Linguistics – Spring 2020

Working Term Paper

**Table of Contents**

[INTRODUCTION 3](#_Toc40408901)

[LOANWORD PHONOLOGY 4](#_Toc40408902)

[A PRELUDE INTO HISTORY 4](#_Toc40408903)

[LITERATURE REVIEW 4](#_Toc40408904)

[THE FACTS SUMMARIZED 5](#_Toc40408905)

[Hanoi Vietnamese Phonology 5](#_Toc40408906)

[French Phonology 7](#_Toc40408907)

[Adaptation Patterns 8](#_Toc40408908)

[FINITE STATE PHONOLOGY 11](#_Toc40408909)

[OVERVIEW OF FINITE STATE PHONOLOGY 11](#_Toc40408910)

[ABOUT THE FOMA TOOLKIT 12](#_Toc40408911)

[Defining phonological natural classes 12](#_Toc40408912)

[Defining phonological rules 12](#_Toc40408913)

[IMPLEMENTATION 13](#_Toc40408914)

[EVALUATION 15](#_Toc40408915)

[CONCLUSION 17](#_Toc40408916)

[References 18](#_Toc40408917)

INTRODUCTION

The existence of foreign loanwords often arises out of language contact due to historical and social factors between groups of peoples. Even the major languages of the world, spoken by a large number of people and whose dominant status testifies for their speakers' success in conquests or their nations' geopolitical importance, have themselves borrowed thousands of words from other tongues: English from Norman French after the Norman conquest of England in 1066, Mandarin Chinese from the northern languages including Mongolian and Manchu, Romance languages from Latin, to name a few. The process of borrowing involves adapting the sounds from the source language to a template of the recipient language, resulting in patterns and rules that reveal interesting facts about the phonology of the recipient language. What is also fascinating is cases where the phonology of the source language influencing the recipient language to the point where foreign sounds are accepted and surface despite their novelty or violations of constraints in the recipient language's phonology.

This paper reports preliminary findings of an effort to model loanword adaptation from French to Vietnamese using a finite-state transducer. The phonological adaptation patterns have been observed and described in the literature to some extent, but no study has evaluated how well the proposed rules work on a corpus of novel words. Previous studies were also fragmented in their descriptions of French-Vietnamese loanword phonology, as different papers focused on accounting for specific adaptation patterns. This study endeavors to unify the proposed adaptation patterns in the implementation of the transducer. A model based on the finite state transducer offers an alternative to traditional phonological description and small-scaled corpus analysis, as it can be rapidly tested, modified, and scaled to apply to millions of words, thus enabling a thorough and precise investigation of how well the rules can be generalized to new word instances. A finite-state transducer also offers a computationally efficient method to capture the regularities of mappings between sounds between the two languages, as its theoretical foundation rests upon operations of relations. Preliminary findings showed that about 48% of a gold test corpus was captured by the finite state transducer rules, thus suggesting that many loanword adaptations were not faithful to the underlying phonology of the source language, likely due to the influence of orthography and a transmission due to textual readings of the words instead of true knowledge about spoken French.

The paper is divided into four main sections: (1) the Loanword Phonology section focusses on loanword phonology of French words in Vietnamese, including a brief prelude about loanword history in Vietnam, leading to a discussion about adaptation patterns of French loanwords in particular. This section also reviews and summarizes results of related studies done on this topic, (2) the Finite State Phonology section describes the main concepts behind applying finite state machine models to phonology, and presents the foma toolkit used to implement the finite state network, (3) the Evaluation section presents the result as the transducer is applied to test corpora and interprets the patterns found in the test corpora, and (5) the Conclusion concludes the paper.

LOANWORD PHONOLOGY

## A PRELUDE INTO HISTORY

Vietnamese is no exception to the loanword phenomenon. An Austroasiatic language spoken by more than 80 million people primarily in Vietnam and in overseas communities where Vietnamese people reside, the language experienced different waves of borrowings in its history.

(Alves, 2009) reported the main borrowing phases in the Vietnamese language, corresponding to periods of extensive foreign influence, in a chapter about this topic in Loanwords in the World's Languages. Before the pre-Han dynasty, it was speculated that Vietnamese borrowed from Proto-Tai. Definitive conclusions about borrowing could be made about the 1000-year period of Chinese domination, where two main borrowing waves were reported. During the first two centuries CE, as the Han dynasty established its political and cultural systems in Vietnam, initial Sino-Vietnamese contact created a stock of Sino-Vietnamese words that have since been regarded as native by the modern Vietnamese people. In the seventh to tenth century, the prosperous and powerful Tang dynasty exerted enormous cultural influence in the Sino sphere, including the popularization of Chinese rhyming dictionaries in East Asia and Vietnam, leading to an influx of Middle Chinese words that were later known as Sino-Vietnamese words. The modern era also witnessed the borrowing of Sino-neologisms coined in China and Japan to describe newly introduced Western concepts. As much as a quarter of Vietnamese vocabulary can be said to be Chinese in origin, and as much as 90% of the loanwords in the database (Alves, 2009) analyzed was from Chinese.

While Chinese has left a decisive mark on the Vietnamese lexicon, French borrowing was not able to gain such strong foothold, for the French colonial period in Vietnam (1898 - 1954) was substantially shorter. The terms borrowed during this period belong to categories such as food and drink, clothing and grooming, agriculture and vegetation, and concepts about the modern world. Most recently, the advent of the Internet became the vehicle of English popularity, and terms related to information technology were often directly borrowed from English. However, the extent to which these terms would stay and become ingrained in the lexicon could only be verified with the passage of time.

## LITERATURE REVIEW

One of the earliest studies done on French loanwords in Vietnamese is Barker's *Phonological Adaptation of French Loanwords in Vietnamese* paper (1969). This study presents six general rules about adaptation from French to Vietnamese, including transformation of nasal vowels, truncation of polysyllabic words, reduction and epenthesis in consonant clusters, adaptation of coda consonants, change in vowel qualities, and the predominance of the level tone in the loanwords. The paper also includes an appendix of 130 loanwords. (Vuong, 2011) work, based on his own thesis (Vuong 1992), describes the dialectal variation of French loans both in the Northern and Southern Vietnam.

More recent studies include a deep analysis into vowels adaptation in *French loanwords in Vietnamese: the role of input language phonotactics and contrast in loanword adaptation* by (Kang et al., 2014), an OT-based account of consonant clusters by (Nguyen and Dutta, 2017), and a corpus-based study on tone assignment and consonant adaptation in *The integration of French loanwords into Vietnamese: a corpus-based analysis of tonal, syllabic and segmental aspects* by (Scholvin and Meinschaefer, 2018). Scholvin and Meinschaefer postulated that due to a low degree of French-Vietnamese bilingualism seen in the Vietnamese population, ‘loanword adaptation has been based on the phonetic surface structure of French, without interference from any knowledge of French phonology’. (Kang et al., 2014) also observed that adaptation was sometimes driven by the orthography, suggesting that some adaptation did not tap into the underlying phonology of the source language but was entirely influenced by the textual readings of words. (Kang et al., 2014) study examined over 1000 French loanwords; however, this corpus was not published.

## THE FACTS SUMMARIZED

Hanoi Vietnamese Phonology

This section provides a sweeping overview of (Hanoi) Vietnamese. Even though Vietnamese has three main regional dialects, the current study is limited to adaptation of French loans to Hanoi Vietnamese dialect only. Therefore, the below phonology tables are directly taken from (Kirby, 2011) work on (Hanoi) Vietnamese phonetics.

A screenshot of a cell phone

Description automatically generated

Figure 1 Hanoi Vietnamese Consonants Inventory

Notable facts about Vietnamese consonants are the presence of the labial voiced implosive, the dental voiced implosive, and the contrast of the dental plosives by aspiration. Phonologically, labialization is reported among the velar stop codas following back rounded vowels /u o ɔ/, and transcriptions in (Kirby, 2011) present the resulting doubly articulated labial velars as k͡p and ŋ͡m. In this paper's implementation, however, these labial velars shall be treated as allophones of /k/ and /ŋ/, respectively, and the Vietnamese IPA transcriptions shall not include k͡p and ŋ͡m in the finite state transducer. Phonotactics of the coda position in Vietnamese also licenses only a subset of consonants to be valid codas in the language: the voiceless plosives /p t k/, the nasals /m n ŋ/, and the approximants /w j/. The set of valid Vietnamese consonant symbols declared in the transducer implementation includes [ ɓ | t | tʰ | ɗ | t͡ɕ | k | ʔ | m | n | ɲ | ŋ | f | v | s | z | x | ɣ | h | w | j | l ].

The below table shows the vowel inventory of Hanoi Vietnamese, consisting of nine vowels and three diphthongs.

A picture containing photo

Description automatically generated

Figure 2 Hanoi Vietnamese Vowels Chart

The vowels /a/ and /ɤ/ have a shortened counterpart in closed syllables, where they are transcribed as /ă/ and /ɤ̆/ respectively. These short vowels can never appear in open syllables. Based on this analysis and Kirby's vowel chart, the set of valid Vietnamese vowel symbols declared in the transducer implementation includes [ i | e | ɛ | ɤ | ɤ̆ | a | ă | u | o | ɔ | ɯ | iə | uə | ɯə ].

Tones in Vietnamese have traditionally been analyzed as consisting of six contrastive tones. The following figure shows the pitch patterns of six Northern Vietnamese tones, as spoken by a female native speaker on the syllable *ma*. The tones, from left to right, are mid-level (ML), mid rising (MR), low falling with breathiness (LF), mid falling rising (MFR), mid rising with creakiness (MRC), and mid falling with creakiness (MFC). Voice quality changes such as breathiness and creakiness are distinctive features of Vietnamese tones.



Figure 3 Pitch patterns of six Northern Vietnamese tones

In the Central and Southern Vietnamese varieties, the MFR and MRC tones have reportedly been merged into one tone, which could be a natural sequence of these tones being similar in contour coupled with the fact that syllables with the MRC tone occur much less often in the language to begin with. Furthermore, the distribution of tones is constrained by the syllable shape for closed syllables: in closed syllables with voiceless stop codas, only the MR and MFC tones are licensed. (Kirby, 2011) divided Hanoi Vietnamese tones into eight contrastive tones, the MR and MFC tones in closed syllables make up new contrastive categories. The below figure illustrated his analysis.

A close up of a map

Description automatically generated

Figure 4 Hanoi Vietnamese Tones

For simplification of the implementation and also following the corpus accompanying (Scholvin and Meinschaefer, 2018) paper, this study follows the traditional analysis of six contrastive tones, coding the tones as follows. The distribution of tones in (Scholvin and Meinschaefer, 2018) corpus shows an overwhelming presence of tone 1, followed by tone 5 and tone 2. The 4, 3, and 6 tones all are very uncommon in loanwords.

|  |  |  |
| --- | --- | --- |
| Tone name | In orthography | Code in foma |
| ML | Ma | 1 |
| MR | Má | 5 |
| LF | Mà | 2 |
| MFR | Mả | 4 |
| MRC | Mã | 3 |
| MFC | Mạ | 6 |

Figure 5 Tone coding in foma

French Phonology

The table below, taken from (Scholvin and Meinschaefer, 2018), illustrates French consonants inventory. It is clear that Hanoi Vietnamese is lacking in the fricative category compared to French, as sounds such as /ʃ/, /ʒ/, and /ʁ/ are not available in the Vietnamese consonants inventory. The set of valid French consonant symbols declared in the transducer implementation includes [ p | b | t | d | k | g | m | n | ɲ | ŋ | f | v | s | z | ʃ | ʒ | ʁ | w | l | j | ɥ].

A picture containing hanging, room, white, clock

Description automatically generated

Figure 6 French Consonants Inventory

The table below shows the set of French vowels, where the nasal vowels are notably distinct from what are considered to be normal vowel sounds in the Vietnamese inventory. Based off of this set of French vowels, the set of valid French vowel symbols declared to be accepted by the transducer includes [ i | y | e | ø | ɛ | œ | ə | a | u | o | ɔ | ɑ | ɛ̃ | œ̃ | ɔ̃ | ɑ̃ ].

A picture containing clock

Description automatically generated

Figure 7 French Vowels Inventory

Adaptation Patterns

Pulling from multiple threads of descriptions about adaptation patterns, this section briefly summarizes the major patterns, specifically organized in a way that would render the rules transparent for a finite state implementation. Minor rules might not be mentioned due to interest of space and time. The main mapping rules are the onset consonant mapping rules, the coda consonant mapping rules, the vowel mapping rules, the tone assignment rules, and the context-free mapping rules.

Onset Mapping Rules

Adaptation of onset consonants frequently involves epenthesis or deletion if the onset consonant is a part of clusters. (Scholvin and Meinschaefer, 2018) describe these patterns in detail. In general, onset singles are replaced by native phoneme whenever discrepancy occurs, such as by adapting p to ɓ, g to ɣ, ʃ to s, ʒ to z, ʁ to z, and j to z. When deletion is applied, the main strategy is for the first consonant to be deleted. The second consonant might be deleted when the second consonant is the liquid /l/. Epenthesis is a much less frequently employed strategy. When epenthesis occurs, the most common vowel to be inserted is /ɤ/ and the tone assigned on the new syllable could be either tone 1 or 2.

Coda Mapping Rules

Coda mappings have to satisfy coda restrictions in Vietnamese, such that only the voiceless stops, nasal, and approximants could function as codas in the surface form. Coda clusters are often repaired by deletion, where the second consonant is usually deleted. The only exception is when the loanword ends with a liquid followed by an /m/. In this case the /m/ would be preserved. The uvular fricative is often integrated into the previous vowel if found in the coda position.

Vowel Mapping Rules

/i/ and /u/ are mapped straightforwardly to their corresponding Vietnamese vowels while other vowels may require some modifications. (Kang et al., 2016) provided very thorough analyses of patterns of vowels adaption, which were summarized in the chart below.

A screenshot of a cell phone screen with text

Description automatically generated

Figure 8 Vowel adaptation, proposed by (Kang et al., 2016)

(Kang et al. 2016) also observed that both French and Vietnamese show contrast between mid-tense and mid lax vowels. Before /ʁ/, adaptation is invariable. In other contexts, adaptation shows a tendency of Loi de position effect, (in other words, lax vowel adaptation is more likely if the vowel ends up in a closed syllable than in an open syllable in the Vietnamese output). The paper also describes the orthographic influence on gemination of consonants and its corollary effect on Loi de position’s application.

Tone Assignment Rules

As French is a stress accent language (Hyman, 2006) while Vietnamese is a tonal language with obligatory tone assignment on every syllable, the loanword adaptation process necessarily involves tone assignment. Two major points about tone assignments were observed by Scholvin and Meinschaefer, firstly most syllables are assigned the mid-level tone (1) and if a closed syllable ends in a voiceless stop, it will be assigned either tone 5 or 6. This is in line with Vietnamese tonal constraint, as mentioned in the previous section. While Scholvin and Meinschaefer noted that 92% of syllables in their corpus ending in a voiceless stop is assigned tone 5, and the remaining 8% is assigned tone 6, it could be argued that this asymmetry is due to the mixture of alternative pronunciations from the Southern dialect, where assignment of tone 6 is preferred. For the purpose of this paper, those words were removed from the corpus, to keep the model consistent with the intended scope of this study. Tone 2 surfaces in a few cases, or six percent in Scholvin and Meinschaefer’s corpus.

Context-free Mapping Rules

Context free mapping rules in the case of French loans in Vietnamese involve transformation that do not depend on the context. All of these involve consonant changes. The mappings of /b/ to /ɓ/, /d/ to /ɗ/, /g/ to /ɣ/, /tʁ/ to / t͡ɕ/, / dʒ/ to /z/, /ʃ/ to /s/, and /ʒ/ to /z/ were judged to be context-free, based on evidence in the test data.

FINITE STATE PHONOLOGY

## OVERVIEW OF FINITE STATE PHONOLOGY

Finite state machines are abstract models representing states and transitions between states. A state diagram is also analogous to a directed graph, or a network in another terminology. Regular languages are formally defined as sets of strings accepted by a finite state machine and can be expressed with a regular expression (Partee et al., 1990). For instance, below is a regular expression written in foma, denoting a regular language and consisting of a set of strings S = {aba, abba, abbba, abbbba, …}, where b can appear one or more times between a’s. The resulting directed graph shows how the acceptor starts from state 0, accepts a and moves to state 1, accepts b, and moves to state 2, where multiple b’s can be accepted while the machine stays at state 2, finally, the acceptor accepts a and moves to the final state 3. This acceptor will not recognize strings that are outside of their set of accepted strings or strings whose individual string does not lead to the next accepting state. This representation also ensures that at least one b is needed between two a’s for the string to be accepted by the finite state acceptor.

define S [ a b+ a ];

A close up of a clock

Description automatically generated

Figure 9 The state diagram for [ a b+ a ]

As regular languages and finite state machines can be represented by sets, set theoretic operations such as union, intersection, subtraction, complementation, apply to them elegantly. Thus, phonological natural classes can be represented by a regular expression, simply by listing the symbols that are allowed to be defined as belonging to a certain class. Certain classes could be written as union or intersection of existing classes, allowing the relationships between classes to be transparent.

Phonological transformations can be likewise modeled by a finite state transducer, developed by Kaplan and Kay in the 1980s (Kaplan and Kay, 1981; Kaplan and Kay, 1994). A finite state transducer models a relation between an upper language and a lower language, whereby both are represented by regular expressions. For example, the figure shown below is the state diagram of a simple replacement rule [ a -> b], where a in the upper language is related to b by the notation <a:b>, b is mapped to itself (or <b:b>), and @ denotes all the other symbols other than a and b, also mapped to itself (or <@:@>). In the context of this study, the upper language consists of French phonemes, syllable boundary marker and the lower language consists of Vietnamese phonemes, syllable boundary marker and tones. Phonological rewrites can be represented by replacement rules, some context-sensitive and some context-free. Because transducers model relations, combining them require composition of multiple machines together in a cascade, in which the order of application was of utmost importance, for the output of a previous machine would function as the input of the next machine in a composition operation.

A drawing of a person

Description automatically generated

Figure 10 State diagram for the simple replacement machine a -> b

## ABOUT THE FOMA TOOLKIT

Foma (Hulden, 2009) is a compiler and programming language that could be used to build finite state automata such as acceptors and transducers outlined above. It is particularly suited for morphological analysis and has support to model various morphological phenomena such as lexical development, dependencies, non-concatenative morphotactics, and so on. It also has many replacement operations that are well-suited to model loanword phonology. The syntax of foma is relatively simple to work with, as the following demonstrates.

Defining phonological natural classes

|  |
| --- |
|  |
|  | define viVowels [ i | e | ɛ | ɤ | ɤ̆ | a | ă | u | o | ɔ | ɯ | "iə" | "uə" | "ɯə" ]; |
|  | define viCons [ ɓ | t | tʰ | ɗ | t͡ɕ | k | ʔ | m | n | ɲ | ŋ | f | v | s | z | x | ɣ | h | w | j | l ]; |
|  | define viSounds [ viVowels | viCons ]; |

Defining phonological rules

Phonological rewrite rules have a basic standard template, shown below.

LHS -> RHS || LC \_ RC

Various alternations are possible to reinforce more sophisticated contexts, such as by using // to indicate that the left context holds on the output side while the right context holds on the input side. The replacement arrow can be enclosed in parenthesis to indicate optional replacement or inversed to indicate inverse replacement direction. In this study, the basic template was used to account for the adaptation patterns in the test dataset.

|  |
| --- |
| Example rule 1: French uvular fricative becomes /z/ in onset position, defined as appearing after word boundary, or after a syllable boundary.  define rChange [ ʁ -> z || .#. \_ , "." \_ ]; # Onset only.  Example rule 2: French voiced bilabial plosive becomes /ɓ/ in any context. |
| define bChange [ b -> ɓ ];  Example rule 3: Assigning the rising tone 5 for syllables ending with /p, t, k/. |
| define RisingTone [ [..] -> 5 || [p | t | k] \_ .#. , [p | t | k] \_ "." ];  Example rule 4: Implementing tone assignment by composing the RisingTone rule with the MidLevel rule. The MidLevel tone was ranked lower than the RisingTone because it is the default rule with freer context. Ranking it higher than the RisingTone might create wrong applications due to the greedy nature of the rule.  define ToneChange [ RisingTone .o. MidLevel ]; |

## IMPLEMENTATION

The test data for evaluation was curated manually based on the online corpus accompanying (Scholvin and Meinschaefer, 2018) paper, which consists of 533 words purported to be still in use in modern Vietnamese, with transcriptions of the original French pronunciations and corresponding Vietnamese pronunciations[[1]](#footnote-1). This corpus contains four columns of data: the Vietnamese transliteration, Vietnamese pronunciation in IPA, the French orthography, and French pronunciation in IPA. When multiple transliterations and pronunciations are possible, different options were separated by commas. The data were originally compiled from (Barker, 1969), Huynh (2008, 2010), and (Nguyen, V. K., 2013), and later expanded via interviews with native speaker informants.

Each pronunciation pair was reviewed by the researcher and certain pronunciations rejected if deemed out of scope, for example pronunciations based on the Southern dialect were pruned. In very few rare cases, a word pair may be rejected if the loan pronunciation bears no resemblance to neither the source language’s pronunciation nor the source language’s orthography. An example of this was ‘ben, ka2 nɔŋ͡m1, benne, bɛn’, which seems to be a mistake in the data. Examples with the foreign phoneme /dʒ/, such as ‘jazz’ and ‘jeep’ were also removed, as they are English words.

The transcription was normalized to be compatible with the convention set up by the study and inconsistent transcriptions were also reconciled. The original corpus uses a mixture of phonemes that French uvular fricative could be mapped to, including /r/ and the flap sound. However, /r/ is very rare in Hanoi Vietnamese, and even though some might pronounce it for foreign words, /z/ is a much more common alternative. Doubly articulated labial velars were also re-written as /k/ and /ŋ/. Syllable boundary was added to the French data by cross referencing with a French dictionary[[2]](#footnote-2) that contains syllable markings. The data format was also normalized such that every pronunciation pair appeared on a line. If multiple pronunciations correspond to one French pronunciation, multiple lines showing the pairs were created. The gold test set after the manual review consists of 608 word pairs[[3]](#footnote-3).

The foma file was developed iteratively and concurrently with the testing phase. In total, about 55 rules were written for the loanword phonology grammar. Files were created of new pronunciation pairs generated by the transducer, of pronunciation pairs in the test corpus but was not accounted for by the transducer, and of pronunciation pairs in the test corpus and was accurately generated by the transducer. Each testing round consists of a review of generated files, which might prompt revision of the test data or of the rules. In total, 48% of the test corpus was accounted for by the model[[4]](#footnote-4). We describe the reasons for the low performance of the model in the next section.

EVALUATION

As observed in previous studies, French loanwords in Vietnamese have pronunciations that are largely influenced by the orthography. The following word pairs are instances of orthography-inspired readings seen in the test set. For example, even though the initial ‘h’ letter is silent in French, it would be often pronounced in the loanwords. The geminate consonants in orthography also create a geminate consonant in the loanword pronunciation, even though the source language does not exhibit this gemination. The final ‘t’ in ‘robot’ is silent in the source language’s pronunciation but surfaces in the loanword pronunciation. These examples show that it would be difficult to achieve high accuracy of loanword pronunciation using a transducer based on the source language’s pronunciations alone.

|  |  |  |
| --- | --- | --- |
| French pronunciation | Loanword pronunciation | Orthography |
| e.ʁɔ.in | he1.zo1.in1 | héroin |
| ɔʁ.mɔn | hɔk5.mɔn1 | hormone |
| se.mi.nɛʁ | se1.mi1.na1 | séminaire |
| sɔ.nɛ | sɔn1.ne1 | sonnet |
| te.nis | tɛn1.nit5 | tennis |
| e.ljɔm | hɛ1.li1 | helium |
| ʁo.bo | zo1.ɓɤt5 | robot |
| ʁo.bo | zo1.ɓot5 | robot |

Figure 11 Orthography-inspired pronunciations

Truncation is another phenomenon that poses difficulty for the model. As Vietnamese is largely made up of monosyllabic and bisyllabic words, lengthy loanwords are frequently truncated, and there seems to be no regular rules governing how this process takes place. The following table showcases cases of truncation seen in the test set.

|  |  |
| --- | --- |
| a.le.zwa | zwa1 |
| al.kɔl | kon2 |
| a.ly.mi.njɔm | ɲom1 |
| bɑ̃.daʒ | ɓăŋ1 |
| o.to.mɔ.bil | ʔo1.to1 |
| sɛʁ.vœʁ o.deʁ ʁə.sə.vœʁ | sɤ1.vɤ1 |
| si.ʁaʒ | si1 |
| si.klo.pus | sik5.lo1 |
| si.ne.ma | si1.ne1 |
| tuʁ.nə.vi | vit5 |
| de.ma.ʁœʁ | ɗe2 |
| man.gu.stɑ̃ | măŋ1.kut5 |
| ma.njez.jɔm | ma1.ɲe1 |
| ma.njez.jɔm | ma1.zi1 |
| ni.tʁo.ʒɛn | ni1.to1 |
| pɔʁt.ba.gaʒ | ɓa1.ɣa1 |
| sa.gu.tje | sa1.ɣu1 |
| su.tjɛ̃.gɔʁʒ | su1.tɕiəŋ1 |
| tuʁ.nə.vi | tuə1.vit5 |
| ɑ̃.ti.gɔn | ti1.ɣon1 |
| ɑ̃.və.lɔp | lop5 |
| ɔk.si.ʒɛn | ʔo1.si1 |
| ka.ʁa.bin | kak6.ɓin1 |
| mo.to.si.klɛt | mo1.to1 |
| va.gɔ.nɛ | ɣɔŋ1 |
| va.gɔ.nɛ | ɣɔŋ2 |
| vjo.lɔ̃.sɛl | sɛn1.lo1 |
| le.gym | ɣim1 |
| li.tɔ.gʁa.fi | li1.to1 |
| li.tiɔm | li1.ti1 |
| lo.ga.ʁit.mə | lo1.ɣa1 |
| lo.ga.ʁit.mə | lo1.ɣa1.zit5 |
| tʁɑ̃.sbɔʁ.dœʁ | tăŋ1.ɓɔ1 |
| tʁɑ̃.sfɔʁ.ma.tœʁ | tăŋ1.fo1 |
| tɑ̃.ʒɑ̃t | taŋ1 |

Figure 12 Examples of truncation

The phenomenon of orthographic influence reflects low degree of bilingualism. The limited number of French loans coupled with a relatively high incidence of irregularity (at least 10% of the test set shows irregularity) calls into question the benefit of using a test set of established loans as a yard stick for measuring the finite state transducer’s performance. The next step of this project would be applying the model on novel words and elicits judgements from bilingual speakers to measure whether or not the generated pronunciations are acceptable. Applying the model on a wider range of words and cross-checked the model with a higher number of speakers would give more hint into how accurate the adaptations are. Another issue that should be checked is over-generation of pronunciations due to loosening of rule restrictions, as the model should strike a balance between accuracy and over accommodation of all peculiarities of the facts.

CONCLUSION

This paper presents preliminary findings of modeling French-Vietnamese loanword phonology with a finite state transducer according to the adaptation patterns reported in the literature. The foma toolkit was used to replicate the rewrite rules that show mappings between sound pairs between the source and recipient language. The result shows that many loanword adaptations were not faithful to the underlying phonology of the source language, likely due to the influence of orthography and a transmission type due to textual readings of the words instead of true knowledge about spoken French. The current accuracy for the model is 47 percent. Truncation is motivated by Vietnamese syllable structure, and it creates many irregularities that make it difficult for a finite state transducer to model well on a test set of established French loans. Future directions for the project include testing the model on a larger test set and consulting opinions of a wider group of native speakers to validate the generated pronunciations on novel words.

References

Barker, Milton E. 1969. The Phonological Adaption of French Loanwords in Vietnamese. *Mon-Khmer Studies Journal* 3:138–47.

Bruce Hayes. 2011. *Introductory Phonology*. John Wiley & Sons.

Huynh, Sabine. 2008. L’assimilation Des mots d’emprunts français à la langue vietnamienne: La question des tons [The Assimilation of French Loanwords into the Vietnamese Language: The Question of Tones]. *Cahiers de Linguistique – Asie Orientale* 37(2):223–40.

Huynh, Sabine. 2010. *Les mécanismes d’intégration des mots d’emprunt français en vietnamien*. Paris: Harmattan.

Kang, Yoonjung, Andrea Hoa Pham, and Benjamin Storme. 2016. French Loanwords in Vietnamese: The Role of Input Language Phonotactics and Contrast in Loanword Adaptation. *Annual Meeting on Phonology* 1–12.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA.

Kirby, James. 2011. Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41(03):381–92.

Lauri Karttunen. 1998. The proper treatment of optimality theory in computational phonology. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing (FSMNLP)*.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th conference of the European Chapter of the Association for Computational Linguistics,* pages 29–32.

Nguyễn, Văn Khang. 2013. *Từ Ngoại Lai trong tiếng Việt*. Thành Phố Hồ Chí Minh: Nhà xuất bản Tổng hợp Thành phố Hồ Chí Minh.

Partee, Barbara & Meulen, Alice & Wall, Robert. (1990). *Mathematical Methods in Linguistics*. 10.1007/978-94-009-2213-6.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. Computational. Linguist. 20, 3 (September 1994), 331–378.

Vương, Toàn. 1992. *Từ gốc Pháp trong tiếng Việt*. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

Vương, Toàn. 2011. *Tiếng Việt trong tiếp xúc ngôn ngữ từ giữa thế kỷ XX. Le vietnamien en contact linguistique depuis la deuxième moitié du XXe Siècle*. Hà Nội: Nhà Xuất Bản Dân Trí.

1. This corpus is accessible in pdf at the original link <https://refubium.fu-berlin.de/bitstream/handle/fub188/23231/loan_corpus_french_viet.pdf?sequence=1&isAllowed=y> or in tsv at <https://github.com/purrfectgrape/vi_vn_loanword_phon/blob/master/loan-corpus-533-raw> [↑](#footnote-ref-1)
2. The French dictionary with syllable marking is accessible from the following repository <https://github.com/msavva/transphoner/blob/master/data/README.md> [↑](#footnote-ref-2)
3. This corpus is accessible at <https://github.com/purrfectgrape/vi_vn_loanword_phon/blob/master/fr-vi-gold.tsv> [↑](#footnote-ref-3)
4. The foma grammar file is accessible from the following link: <https://github.com/purrfectgrape/vi_vn_loanword_phon/blob/master/fr-vi.foma> [↑](#footnote-ref-4)