

A finite-state morphological analyzer using foma for Klingon, a constructed language

Matthew David Dreselly Thomas
University of Colorado at Boulder
matt@bocosoft.net

Abstract

This paper presents a morphological analyzer for Klingon, constructed language, developed following the official description of the morphology in the Klingon Dictionary (Okrand, 1992).

The open-source tool *foma* for constructing finite-state automata and transducers is used for building the morphological analyzer.

1 Introduction

Klingon was created by Marc Okrand, an American Linguist, for the 1984 movie *Star Trek III*. Klingon is spoken by the fictional Klingons in the *Star Trek* universe, so Okrand designed the Klingon language to sound alien by using grammatical constructs and sound combinations that are rarely used in natural language (Institute, b; Institute, c).

Klingon has a complete grammar that can be spoken and used as a real language. Marc Okrand formalized the Klingon grammar and vocabulary in *The Klingon Dictionary* (Okrand, 1992). The official writing system using the Latin alphabet and is case-sensitive. The alien alphabet, called *pIqaD*, was created for visual effects in the movies by the Astra Image Corporation.

The number of fluent Klingon speakers is not known, but has been estimated between a few dozen (Okrent, 2009) to around 100 (Kelly, 2013). *The Klingon Dictionary* (Okrand, 1992) has sold more than 300,000 copies and has been translated into 4 languages. Klingon has entered into popular culture, and literary works such as Shakespeare's *Hamlet*, *Gilgamesh*, and "A Christmas Carol" have been translated into Klingon. The current lexicon contains almost 3,000 words (Zrajm,) and only grammar and words coined by Marc Okrand are considered canonical by the Klingon speaking community (Institute, a).

2 FST Implementation

The morphological analyzer FST consists of two files; a *lexc* file and a *foma* file. References to *The Klingon Dictionary* and other sources are captured in comments throughout the FST files for reference and as an explanation of implementation decisions.

2.1 *lexc* file

The *lexc* file describes Klingon morphology and contains the basic set of Klingon lexical entries. I use A-P entries of the KLCP1, which are the 502 basic words of first level of KLI's "Klingon Language Certification Program" (Zrajm,). Each lexical entry contains the stem and the English meaning on the upper-side (analysis) of the FST.

Beesley and Karttunen (Beesley and Karttunen, 2002) flag diacritics are used to restrict the stem and affix combinations according to the grammar description.

2.2 *foma* file

The *foma* file creates an FST from the *lexc* and adds morphophonological rewrite rules. The single Klingon morphophonological rule is implemented in the *foma* file as a rewrite rule.

3 Klingon Morphology

Klingon is an agglutinative language that uses Object-Verb-Subject word order. Okrand categorizes all Klingon words into Nouns, Verbs, and Other. The morphological analyzer is specifically designed and tested to work with Nouns and Verbs, however some additional work was done to account for the words in the 'Other' category.

3.1 Nouns

Nouns can take five types of suffixes. Each suffix type, if present, must be appended to the noun stem in a specific order and there can be no more

than one suffix per type. Nouns are divided into three categories; body-parts, beings capable of language, and general nouns. Some suffixes can only be appended to certain noun categories or their meaning changes when they are appended.

Nouns use the "NOUNTYPE" feature flag to track the noun category; @U.NOUNTYPE.BODYPART@, @U.NOUNTYPE.CAPABLELANGUAGE@, @U.NOUNTYPE.OTHER@ and @U.NOUNTYPE.LOCATION@.

Below are the noun suffix types and the relevant rules for each type.

Suffix Type 1 Augmentative/diminutive

The endearment suffix *-oy* adds a dash '-' which is replaced with either null or with a glottal stop in the *foma* file depending on whether the last letter of the noun stem is a consonant or vowel, respectively.

Suffix Type 2 Number

Noun Type	Suffix
Body-Part	-pu'
Language-Capable	-Du'
Generic	-mey

The *-mey* suffix may also be used with body-part and beings-capable-of-language nouns, but the resulting word acquires an additional meaning "*all over the place*" on top of the plural meaning.

Klingon	English
jup	friend
jupDu'	friends
jupmey	friends all over the place

Table 1: Generic Plural Suffix Usage with Inherently Singular Nouns

Some nouns are inherently singular or inherently plural. For example the Klingon word *peng* means torpedo (singular) but *cha* means torpedos (plural). The singular forms of these types of nouns may also take the *-mey* generic plural suffix, but the resulting word also acquires the "*all over the place*" meaning.

If a noun is inherently singular or inherently plural, it is marked using the Positive flag diacritic @P.SINGULAR.IRREGULAR@ or @P.PLURAL.IRREGULAR@.

Klingon	English
peng	torpedo
pengmey	torpedos all over the place
cha	torpedos
*chamey	<i>Not allowed</i>

Table 2: Generic Plural Suffix Usage with Language-Capable Nouns

Lastly, nouns in Klingon are not required to take any plural suffix in order to be plural. The singular form of a noun may be used where the plural form would normally be used, and this is perfectly grammatical. The meaning is disambiguated using the context of the word in the sentence. Therefore *jup* means both "friend" and "friends" depending on the context. There are no articles in Klingon.

The Type-2 suffix group uses the NOUNTYPE, SINGULAR and PLURAL flag diacritics to control which suffixes can be applied and what their meaning is.

Suffix Type 3 Qualification

Suffix Type 4 Possession/specification

Possession suffixes are selected depending on whether the *object* being referred to is language-capable.

Possessive Suffixes	Generic	Language Capable
my	-wIj	-wI'
your	-IIj	-II'
his, her, its	-Daj	-Daj
our	-maj	-ma'
your (plural)	-raj	-ra'
their	-chaj	-chaj

Language-capable nouns can still take the generic suffix but the meaning becomes derogatory (Okrand, 1992). The noun type is tracked using the NOUNTYPE flag and in this case the tag [DEROGATORY] is appended to the analysis.

Suffix Type 5 Syntactic markers

(Okrand, 1992) notes that the locative suffix *-Daq* cannot be appended to three adverb-like Klingon nouns.

"It is worth noting at this point that the concepts expressed by the English adverbs (here, there), and (everywhere) are expressed by nouns in Klingon: *naDev* (hereabouts), *pa'* (thereabouts),

Dat (everywhere). These words may perhaps be translated more literally as “area around here,” “area over there,” and “all places,” respectively. Unlike other nouns, these three words are never followed by the locative suffix.”

These three nouns are tracked by setting the NOUNTYPE flag (@U.NOUNTYPE.LOCATION"@"). These can be prevented from taking the locative suffix using the Disallow flag test (@D.NOUNTYPE.LOCATION@).

Noun Tag

After all suffixes are appended, one of three options are chosen for the noun tag.

- If the VERBTYPE.STATIVE flag has been set, it means that a stative verb was using the Type-5 Noun suffix as explained in the Verbs section. Therefore no Noun tag needs to be added to the word analysis.
- Otherwise if the MOOD flag has *not* been set, then we can be sure that a noun is being tagged and can thus append the [N] tag to the analysis.
- Lastly, if the MOOD flag *has* been set then we know that the word stem is a verb and we tag the current word with [DVN].

3.2 Verbs

Verb can take both prefixes and suffixes. Prefixes specify both the subject and object of the verb, including whether or not a subject and/or object exists. Verbs can use either the **Indicative mood** or the **Imperative mood**, which can also affect whether some suffixes can be used. There is no infinite form of a verb in Klingon.

Verbs are categorized into **stative** and **dynamic**. Klingon does not use adjectives so stative verbs are used to describe nouns. For example to say “red ship”, one would use the stative verb Doq meaning to be red/orange, “*Duj Doq*”.

Verbs conform to nine ordered suffix types. Just like nouns, each suffix type, if present, must be appended to the verb stem in a specific order. There can be no more than one suffix per type.

Below are the verb suffix types and the relevant rules and flag diacritics for each type. Verbs can take either an Indicative or Imperative mood,

which affects the suffixes they can take and their meaning.

Prefix Indicative or Imperative

The verb mood is tracked using a MOOD flag @U.MOOD.INDICATIVE@, @U.MOOD.IMPERATIVE@ or @U.MOOD.INFINITIVE@. In reality Klingon does not use verb infinitives however I apply this mood to restrict verb prefixes when analyzing verbs used adjectivally, which should not have a prefix.

The Indicative prefixes indicate both the subject (if any) and object (if any) of the verb. Verbs with no object are tracked using the flag @P.OBJECT.NONE@. Verbs with no subject are tracked using the flag @P.INDICATIVE.UNSPECIFIED@. Verb subjects are tracked with the SUBJECT flag (@U.SUBJECT.PLURAL@ or @U.SUBJECT.PLURAL@). If there is no subject for the verb, then this is tracked with @P.INDICATIVE.UNSPECIFIED@.

Finally the verb type (dynamic vs stative) is tracked using the VERBTYPE flag.

Rover Suffixes

In addition, a “rover” suffix type can be appended immediately after the verb stem or between the nine suffix types, however rover suffixes cannot be appended after a Type 9 suffix. There are no restrictions on the number of rover suffixes that may be used. A rover suffix modifies whatever directly immediately precedes it. For example the **negating** rover suffix *be*’ can be appended immediately after the verb stem and/or after any other suffix type.

Klingon	English
Dub	to improve
Dubbe’	to not improve
DubmoH	to cause to improve
DubmoHbe’	to not cause to improve
Dubbe’moH	to cause to not improve
Dubbe’moHbe’	to not cause to not improve

Table 3: Generic Plural Suffix Usage with Language-Capable Nouns

Rover suffixes can even be appended after other rover suffixes, e.g. *Dubbe’be*’ meaning *to not not improve*.

The [REFUSE] tag/suffix *Qo*’ can only

be appended if the MOOD is Indicative (@U.MOOD.INDICATIVE@).

The suffix *be'* meaning "negation" can only be appended if the MOOD is Indicative (@U.MOOD.INDICATIVE@). Otherwise if the MOOD is Imperative, a different suffix *Qo'* must be used to negate the verb. The Indicative negation suffix *Qo'* can only occur at the end of the nine suffix types, so I turn the verb suffix gate type switches on to apply this restriction. If the MOOD is Imperative but the suffix *Qo'* is used, then the suffix takes on a meaning of refusal.

Suffix Type Gates

Just like nouns, verb suffix types need to occur in a certain order. However because the rover suffix type can occur *between* the other suffix types, therefore the morphology needs to track the last appended suffix type. To address this, I employed "suffix type gates". When the input leads through a specific suffix type, for example suffix type 5, the type gate denies entry if suffix type 5 has already been turned-on (@D.VERBTYPE5.ON@). If suffix type 5 flag has not been turned on, then the gate turns it on (@P.VERBTYPE5.ON@) as well as all previous suffix types; @P.VERBTYPE1.ON@ @P.VERBTYPE2.ON@ @P.VERBTYPE3.ON@ @P.VERBTYPE4.ON@. In this way, the FST is restricted from repeating the same suffix type more than once, even after appending a rover suffix.

Suffix Type 1 Reflexive, Reciprocal

This suffix type can only be used with the set of prefixes that mean "no object".

The *-chug* reciprocal suffix is additionally restricted to use when the subject is plural (@U.SUBJECT.PLURAL@) and the verb type is dynamic (@U.VERBTYPE.DYNAMIC@).

Suffix Type 2 Volition, Necessity

Suffix Type 3 Inceptive, Inchoative

Suffix Type 4 Causative

Suffix Type 5 Undefined Subject, Capability

The undefined subject suffix *lu'* can only be and is required to be used if the subject of the verb is not specified (@R.INDICATIVE.UNSPECIFIED@). To make sure that both requirements are satisfied, we clear the INDICATIVE flag after appending the undefined-subject suffix and at the end of the word analysis use a disallow test to make sure that the flag has been cleared

(@D.INDICATIVE.UNSPECIFIED@).

Suffix Type 6 Perfection, Uncertainty

Suffix Type 7 Aspect

The continuous suffix *-taH* set a flag diacritic @P.VERBSUFFIX.CONTINUOUS@ which will be referenced later by a type 9 suffix.

Suffix Type 8 Honorific

Suffix Type 9 Syntactic markers

The person suffix *-wI'* can only be used when the verb has no prefix, so the unification test @U.MOOD.INFINITIVE@ is applied. This suffix makes the verb a regular noun to which all regular noun suffixes can be applied. Therefore instead of continuing onto the verb suffix path, the person suffix reroutes to the NounSuffix.

The simultaneous suffix *-vIS* can only be used when the type 7 continuous suffix *-taH* has been appended, so the require flag test @R.VERBSUFFIX.CONTINUOUS@ is used.

Verb Tag

Verbs are tagged with [V]. if the verb is stative @U.VERBTYPE.STATIVE@ then an additional tag [STATIVE] is appended as well.

If the mood is infinitive @U.MOOD.INFINITIVE@, stative nouns can attach a noun suffix type 5. Otherwise either the [INDICATIVE] or [IMPERATIVE] tag is appended to the verb using the unification tests (@U.MOOD.INDICATIVE@ and @U.MOOD.IMPERATIVE@).

3.3 Other Words

Pronouns

There is a closed set of specific and independent words for pronouns. These pronouns are all accounted for within the *lex* file.

Klingon	English
jIH	to I, me
soH	you
ghaH	he/she, him/her
'oH	it
'e'	that
net	that
maH	we, us
tlhIH	you (plural)
chaH	they, them (Language-Capable)
bIH	they, them (Generic)

Table 4: Klingon Pronouns

Pronouns may be used as nouns however appending suffixes to pronouns is outside the scope of this project.

Numbers

The Klingon number system is base ten and numbers are combined just like in English with higher-order number coming first. For example 10,503 would be ten THOUSAND five HUNDRED and three. The thousands value needs to precede the hundreds value. In order to enforce this ordering, I needed applied the same gating system that was used for the verb ordered suffix types.

Numbers can take the ordinal suffix *-DIch* or Numbers can act as adverbs if given

Conjunctions

In Klingon a different set of conjunctions is used when combining sentences than when combining words.

English	Words	Sentences
and	je	'ej
or	joq	qoj
xor	ghap	pagh
constrast		'ach / 'a

Table 5: Klingon Conjunctions

Adverbials

Adverbials take a single reversal suffix *Ha'* and no prefixes. Numbers can also act as adverbs if they take the repeat-count suffix *-logh*.

Exclamations

There are a closed set of exclamations and they are handled in the *lexc* file. Exclamations do not take affixes.

Names

A few common Klingon names are added to the *lexc* file without any special handling.

Questions

There are a closed set of questions and they are handled in the *lexc* file. Questions do not take affixes.

4 Klingon Morphophonology

Klingon has only a single morphophonological rule. The endearment noun suffix *-oy* is prepended with a glottal stop ' if the noun stem ends with a vowel. This rule is implemented in the *foma* file.

Noun Stem	With Endearment Suffix
jup (friend)	jupoy (dear friend)
cha (torpedo)	cha'oy (dear torpedo)

Table 6: Endearment Suffix

5 Evaluation

More than 150 test cases are contained in a separate *foma* file. These test cases follow Okrand's examples in The Klingon Dictionary. The morphology *lexc* was modified until all test cases passed and recall reached 100%.

In order to obtain a quantitative analysis of the morphological analyzer, I compiled a list of words from The Klingon Dictionary and The Klingon Hamlet (096,). Of the 267 words on the list, the morphological analyzer accounted for 264. The remaining 3 words were *DatDaq*, *tIpuq* and *naDevDaq*. Both *DatDaq* and *naDevDaq* are explicitly forbidden by Marc Okrand in The Klingon Dictionary as there are location nouns with the locative suffix, therefore these words are technically not grammatically. The last word *tIpuq* comes from the Sonnet-18 translation in Hamlet and appears to translate to "bud" (as in rose bud), however this word does not appear in any canonical sources of Klingon. Therefore I consider the morphological analyzer to have achieved 100% recall for the target word list.

References

- Hamlet, Prince of Denmark: The Restored Klingon Version*. Klingon Language Institute.
- Beesley, K. R. and Karttunen, L. (2002). *Finite-State Morphology: Xerox Tools and Techniques*. Xerox Corporation.
- Institute, T. K. L. Canonical Sources. http://www.kli.org/wiki/Canonical_sources. Accessed: 2015-12-15.
- Institute, T. K. L. Development and Use of the Klingon Language. <http://www.kli.org/about-klingon/klingon-history/>. Accessed: 2015-12-15.
- Institute, T. K. L. The Sounds of Klingon. <http://www.speakklingon.info/about-klingon/sounds-of-klingon/>. Accessed: 2015-12-15.
- Kelly, S. M. (2013). How Klingon Became a Universal Language. <http://mashable.com/>

2013/05/18/klingson-star-trek/. Accessed: 2015-12-15.

Okrand, M. (1992). *The Star Trek: The Klingon Dictionary*. Simon and Schuster.

Okrent, A. (2009). *In the Land of Invented Languages: Esperanto Rock Stars, Klingon Poets, Loglan Lovers, and the Mad Dreamers Who Tried to Build A Perfect Language*. Spiegel & Grau.

Zrajm. Klingon Word Database 1st edition.
<http://mashable.com/2013/05/18/klingson-star-trek/>.