**Ville Rissanen**

# Translitterointi japanin kielen syötemetodina

Tietotekniikan kandidaatintutkielma

25. huhtikuuta 2022

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

**Tekijä:** Ville Rissanen

**Yhteystiedot:** ville.rissanen@linux.com

**Ohjaaja:** Antti-Jussi Lakanen

**Työn nimi:** Translitterointi japanin kielen syötemetodina

**Title in English:** Transliteration as an input method for the Japanese language

**Työ:** Kandidaatintutkielma

**Sivumäärä:** 24+0

**Tiivistelmä:** Tutkielmassa esitellään japanin kielen yksityiskohdat, jotka vaikuttavat kielen latinasointiin ja siten myös sen translitterointiin. Näiden perusteella esitellään deterministinen translitteraatiojärjestelmä Jikara, jonka avulla voidaan translitteroida erityistä japaninkielistä latinasoitua syötettä japaninkieliseksi tulosteeksi ja toisinpäin. Lisäksi tutkitaan lyhyesti esitellyn translitteraatiojärjestelmän mahdollista käyttöä tekstin häviättömään pakkaamiseen, mutta translitteroidun tekstin pakkaussuhde ei ole edullinen verrattuna UTF-8 esitystapaan.

**Avainsanat:** translitterointi, japanin kieli, japani

**Abstract:** A brief overview of the japanese language is given with focus on the points that are of interest to romanisation and hence transliteration. Based on these points the thesis presents a deterministic system of transliteration, Jikara, which can be used to transliterate a specific japanese romanised input to native japanese output and vice versa. In addition a short study is performed to see if the transliteration system can be used to achieve lossless archving of text, but the compression ratio of the transliterated text is not favourable compared to the UTF-8 presentation of the text.

**Keywords:** transliteration, Japanese language, Japanese

# Taulukot

# Sisällys

# 1 Johdanto

Japanin kieli koostuu kolmesta erilaisesta kirjainmerkistöstä: kiinasta lainatuista merkeistä ja kahdesta tavustosta. Kiinalaisia merkkejä kielessä on yli 50 000, joista arkikäyttöönkin on määritelty noin pari tuhatta merkkiä (Frellesvig 2010文化庁 2010). Tavustoissa on 45 yleisesti käytettyä tavumerkkiä, mutta koska tavustoja on kaksi jotka kuvaavat täsmälleen samoja äänteitä olisi merkkejä yhteensä 90. Näin ollen vain yksi tavusto mahtuu tavalliselle näppäimistölle.

Japanin kielen kirjoittaminen sähköisillä laitteilla tapahtuu yleisesti käyttämällä niin sanottua Input Method Editor -ohjelmaa, joka ehdottaa merkkejä kun käyttäjä kirjoittaa joko latinalaisia aakkosia tai japanilaisia tavumerkkejä. Käyttäjä valitsee ohjelman ehdottamista merkeistä tai ilmaisuista graafisen käyttöliittymän avulla, jonka sitten laite korvaa kirjoitetun syötetekstin tilalle. Tämä syötetapa ei mahdollista alkuperäisen syötetekstin palauttamista lopullisesta tekstistä, joka on tämän tutkielman esittämän syötetavan etu verrattuna muihin syötetapoihin.

Käänteistä transliteraatiota japanin kielelle sekä muille kielille on kehitetty konekääntämisen näkökulmasta, mutta tutkielmassa esitetty metodi toimii molempiin suuntiin determinisitisesti. Sen lisäksi tutkielmassa esitetty Jikara-syötetapaa voidaan mahdollisesti käyttää japanin kielen häviättömään pakkaamiseen tietyissä olosuhteissa. Konseptin toimivuutta tutkitaan lyhyesti tutkielmassa transliteraatiomenetelmän esittelyn jälkeen.

Koska japanin kielen esittämiseen on useita eri tapoja, esittelen seuravaaksi tässä tutkielmassa käytössä olleet käytännöt japaninkielisen sisällön ja niiden latinaistuksen esittämiseen. Tässä työssä ei käytetä sirkumflekseja tai pituusmerkkejä, vaan kaikki merkit latinisoidaan niin kuin ne kirjoitetaan Hepburn-translitteraatiossa ja kanjien tapauksessa kuten furiganat kirjoitetaan. Tästä syystä kaikkiin kanji-sanamerkkeihin on lisätty merkkien päälle niiden lukutapa furigana-merkeillä. Poikkeuksena tähän sääntöön ovat teemapartikkeli は (ha => wa), illatiivipartikkeli へ (he => e) ja akkusatiivipartikkeli を (wo => o). Tämä mahdollistaa lähes yhtenäisen esitystavan esitellyn

järjestelmän syötteen ja tämän tutkielman ulkoasun välillä, mutta säilyttää likimääräisesti oikean ääntämistavan luettaessa.

# 2 Japanin kielen erityispiirteitä

Ennen japanin kielen translitteroinnin tutkimista on tarpeellista tietää kielen erityispiirteitä, jotka vaikuttavat translitterointiprosessiin. Tässä luvussa käydään läpi olennaisimmat japanin kielen erityispiirteet tutkimuksen näkökulmasta.

## 2.1 Merkistöt

Japanin kielessä käytetään pääasiallisesti kolmea eri merkistöä: hiraganat, katakanat ja kanjit (Henshall, Seeley ja De Groot 2003 Yamaguchi 2007). Näistä kaksi ensimmäistä on tavumerkistöjä ja kanjit ovat kiinan kielestä lainattuja sanamerkkejä. Edellämainittujen lisäksi arabialaisten numeroiden käyttö on suhteellisen yleistä, mutta latinalaisten aakkosten käyttö on yhä suhteellisen harvinaista (Chikamatsu ym. 2000).

Ongelmia translitteroinnin suhteen aiheuttaa se, että hiraganat, katakanat ja kanjit romanisoidaan identtisesti, koska latinalaisten aakkosten rajallisten merkkien takia translitteroidaan kaikki merkit niiden lukutavan mukaisesti. Arabialaiset numerot ja kanjeilla kirjoitetut numeraalit myös luetaan ja translitteroidaan samalla tavalla. Tällöin latinalaiseen aakkostoon translitteroidusta tekstistä on mahdotonta sanoa täydellä varmuudella, mikä alkuperäisen tekstin merkistö on ollut kyseisen sanan tai tavun kohdalla.

$$3 = \text{さん} = \overset{\text{さん}}{\text{酸}} = \text{サン}$$

$$san$$

Kaikki neljä sanaa romanisoidaan: ”*san*”, mutta ensimmäinen on arabialainen numero kolme, toinen on hiraganoilla kirjoitettava nimen jälkeen käytettävä kohtelias puhuttelumuoto, kolmantena oleva kanjimerkki tarkoittaa happoa ja neljäs on englannin kielen lainasana Auringolle tai pyhimykselle katakanoilla kirjoitettuna.

## 2.2 Homofonit

Japanin kielessä esiintyy myös useita samaan merkistöön kuuluvia sanoja ja merkkejä, joilla voi olla eri merkitykset, mutta identtinen lukutapa ja siten myös identtinen romanisointi. Romanisoidusta tekstistä voi päätellä merkkejä kontekstin avulla samalla tavalla kuin puhutussa kielessä, mutta ilman ulkopuolista kontekstia on merkitysten päätteleminen haastavaa.

日(ひ)= 火(ひ)= 灯(ひ)

*hi*

Esimerkiksi kaikki kolme ylläolevaa merkkiä luetaan tavuna "*hi*". Ensimmäinen tarkoittaa päivää, Aurinkoa tai auringonvalo, toinen tarkoittaa tulta ja kolmas valoa, lamppua tai soihtua. Esimerkkien lisäksi on myös muita merkkejä, jotka voidaan lukea "*hi*".

## 2.3 Kanjimerkit

Toisin kuin tavumerkit, kanjimerkeillä on yleensä kaksi tai useampi lukutapa. Nämä jaetaan kiinalaisiin ja japanilaisiin lukutapoihin, mutta merkillä voi olla myös vain yksi lukutapa. Yleisenä nyrkkisääntönä voidaan pitää, että japanilaista lukutapaa käytetään merkin esiintyessä yksin ja kiinalaista lukutapaa, kun merkki on osa useamman kanjin muodostamaa sanaa, mutta tähän sääntöön on poikkeuksia.

日曜日(にちようび)= ***nichi****you****bi***

大火(たいか)= *tai****ka***

電灯(でんとう)= *den****tou***

## 2.4 Geminaatat

Geminaatat, eli pitkät konsonanttiäänteet, japanin kielessä kaksinkertaistavat seuraavan tavun ensimmäisen konsonantin. Geminaattaa voidaan käyttää vain tietyillä konsonanteilla. N-kirjaimen geminaatat kirjoitetaan N-tavumerkillä. Tämä merkitään sokuon (hiragana っ/katakana ッ) merkillä, joka on vain pienikokoinen ”tsu”tavumerkki. Se asetetaan ennen toistettavan konsonantin tavua ja sillä ei ole muuta vaikutusta lausumiseen. Sokuonia voidaan käyttää myös glottaaliklusiilin merkitsemiseen puhutun kielen kirjoituksessa puhujan ollessa yllättynyt tai vihainen.

待(ま)って = *matte*

どんな = *donna*

## 2.5 Kaksoisvokaalit

Kaksoisvokaalit kirjoitetaan eri tavoin riippuen kumpaa tavumerkistöä käytetään. Hiragana-tavumerkeillä kaksoivokaalit tuotetaan kirjoittamalla toinen samanlainen vokaali edellisen samaan vokaaliin päättyvän tavun perään. Katakana-tavumerkeillä toistuvat vokaalit merkitään chouonpu-vokaalinpidennysmerkillä (ー). Chouounputa voidaan käyttää myös epävirallisesti hiraganatavuilla.

さあ = *saa*

コーヒー = *koohii*

## 2.6 Digrafit

Japanin kielessä esiintyy myös kahden tavumerkin yhdistelmiä, jotka muodostavat oman äänteensä. Näissä digrafeissa ensimmäinen merkki on kirjoitettu normaalisti ja seuraava merkki on kirjoitettu pienenä. Näiden digrafien latinaistus on kahdesta kolmeen merkkiä pitkä.

$$しゃ = sha$$

$$じょ = jo$$

## 2.7 Diakriittiset merkit

Tavumerkkien äännettä voidaa pehmentää tai koventaa lisäämällä tavumerkkiin diakriittinen merkki. Tällöin tavumerkin ensimmäinen tai ensimmäiset konsonantit lausutaan eri tavalla ja viimeinen vokaali samalla tavalla kuin normaalisti.

$$ば = ba$$

$$ぱ = pa$$

## 2.8 Partikkelit

Jotkin tavumerkit luetaan eri tavalla kuin normaalisti, kun ne toimivat partikkeleina virkkeessä. Näitä ovat ovat は, を ja へ. Toimiessaan partikkeleina translitteroidaan tavut seuraavasti:

- は (ha) = ”wa”
- へ (he) = ”e”
- を (wo) = ”o”

Kuitenkin näillä vastaavilla ääntеillä on omat tavumerkkinsä わ (wa), お (o), え (e), joten translitteroidusta tekstistä pitää kontekstin avulla päätellä kumpaa tavumerkkiä tarkoitetaan kontekstin perusteella.

## 2.9 Iteraatiomerkit

Toistuvat merkit voidaan myös kirjoittaa käyttämällä iteraatiomerkkejä ensimmäisen kirjoitetun kanji-, hiragana tai katakanamerkin jälkeen. Jokaisella merkistöllä on oma iteraatiomerkkinsä: kanjeille 々, hiragananoille ゝ ja katakananoille ヽ, joista on myös

tavumerkeille omat dakuten-merkityt versiot, jotka muuttavat toistetun tavun ääntämistä. Kanjien iteraatiomerkki voidaan ääntää eri tavalla kuin alkuperäinen merkki tai sitten ei. Tavumerkkien iteraatioita ei käytetä yleisesti jolleivat ne kuulu erisnimeen, mutta kanjien iteraatiomerkkiä käytetään yleisesti.

何々(なになに) = *naninani*

日々(ひび) = *hibi*

じゞ = *jiji*

ロヽノア = *roronoa*

## 2.10 Kirjoitusasu

Kaikkia edellämainittuja seikkoja translitteroinnin kannalta vaikeuttaa japanin kielen kirjoitusasu. Kirjoitetussa tekstissä ei käytetä välejä sanojen välillä, joka tekee automaattisesta sanojen rajojen löytämisestä haastavaa. Kanjimerkeillä kirjoitettavia sanoja voidaan käyttää myös hiraganoilla tai katakanoilla kirjoitettuna lyhyissä teksteissä kuten tekstiviesteissä tai sävytyskeinona, mutta se ei vaikuta sanojen romanisoituun muotoon. Kirjoitetussa kielessä ei ole myöskään erityistä tapaa merkitä erisnimiä, mutta erisnimillä on oma lukutapansa, vaikka ne kirjoitetaan samoilla merkeillä kuin normaalit sanat.

田中(たなか)さんがコーヒーを飲(の)みます。

*Tanaka-san ga koohii o nomimasu.*

# 3 Syötemetodit ja translitteraatiojärjestelmät

Tässä luvussa käydään läpi kuinka translitterointi ja syöte tapahtuu olemassaolevilla järjestelmillä.

## 3.1 Translitteraatiojärjestelmät

Japanin kielen translitterointiin on kehitetty useita eri järjestelmiä, jotka painottavat kielen eri osa-alueita (Eells 1952). Näistä tunnetuin on Hepburn-translitteraatio, joka myös käytetyin. Muita järjestelmiä ovat ISO-standardoitu Kunrei-shiki, sen edeltäjä Nihon-shiki ja JSL. Suomen kielen lautakunta antoi vuonna 1957 suosituksen, että suomenkielisen tekstin yhteydessä tulisi käyttää suomeen mukautettua Hepburn-translitteraatiota. Suomen suurin sanomalehti, Helsingin Sanomat, on 2000-luvun alkupuolella alkanut käyttämään englantilaistyylistä Hepburn-latinaistusta, josta puuttuu vokaalien pituusmerkit (Itkonen ja Maamies 2002), jota käytetään myös tässä tutkielmassa.

Lisäksi koneellista syötettä varten on kehitetty latinaistusjärjestelmiä kuten ワープロローマ字 (waaporo-romaji), joka tulee englanninkielen sanasta ”word processor”. Waaporo-romajien järjestelmä kehitettiin japanin kielen syöttämiseen tekstieditoreihin ja on käytössä edelleen yhtenä syötemuotona esimerkiksi Input Method Editor -ohjelmistoille. Waaporo-latinaistus oli aiemmin standardoitu, mutta standardi on vanhentunut vuodesta 2010 lähtien ja sitä ei ole uusittu (日本産業標準調査会 2000).

## 3.2 Input Method Editor

Japanin kielen syöttämiseen erinäisiin tietoteknisiin laitteisiin on kehitetty erilaisia menetelmiä. Näistä yleisin on tällä hetkellä Input Method Editor (IME) -ohjelmistot, joille voidaan syöttää merkkejä sanan tai yleisen ilmauksen verran latinalaisilla aakkosilla tai japanilaisilla tavumerkeillä, joista ohjelma muodostaa ehdotuksia todelliseksi syötteeksi. Koska IME:n täytyy pystyä lukemaan kaikki näppäimistön syöte ja korvaamaan

käyttäjän jo kirjoittamaan syötettä on IME yleensä valinnainen osa itse käyttöjärjestelmäohjelmistoa tietoturvasyistä. Yleensä japanilaisissa näppäimistöissä voi vaihtaa latinalaisten aakkosten ja japanilaisten tavumerkkien välillä erityisellä ひらがな/カタカナ -näppäimellä, joka esiintyy japanilaisissa näppäimistöasetteluissa muiden translitteraatioprosessia ohjaavien näppäinten lisäksi.

# 4 Jikara-translitteraatio

Tässä luvussa käydään läpi kuinka Jikara-translitteraatio toimii ja kuinka translitterointi-ohjelmaa voidaan käyttää japanin kielen kirjoittamiseen ASCII-merkistöllä tai japanin kielen tallentamiseen ASCII-merkistöön.

## 4.1 Esittely

Jikara on translitteraatiojärjestelmä, joka pyrkii waapuro-translitteraation mukaisesti takaamaan 1:1 tuottamaan tietylle kirjainyhdistelmälle aina saman merkin. Syötetapa on romaji-syöte, eli tuotettava teksti perustuu ASCII-merkistöön. Toisin kuin muissa syötejärjestelmissä, Jikarassa ei ole merkistötiloja, vaan kaikki teksti voidaan kirjoittaa normaalisti ilman erillistä ohjelmaa ja voidaan sitten translitteroida myöhemmin. Tämä ei kuitenkaan estä reaaliaikaista translitteraatiota. Tavoitteena on tuottaa syntaksi, joka voidaan palauttaa takaisin alkuperäiseen kirjoitusasuunsa. Jikara jakaa nimensä samannimisen translitteraatio-ohjelman kanssa, joka toteuttaa Jikara-translitteraation.

## 4.2 Yleinen syntaksi

Kaikki merkit translitteroidaan niin kuin ne on kirjoitettu. Kanjien tapauksessa tämä tarkoittaa kanjin luentaa, joilla furigana-merkit kirjoitettaisiin. Ohjelman sitä tukiessa voi käyttää haluamaansa tavumerkkien romanisointia, mutta tällä hetkellä ohjelma toteuttaa vain Hepburn-romanisoinnin. Pitkät vokaalit kirjoitetaan toistamalla vokaali. Geminaatat kirjoitetaan toistamalla konsonantti. Pisteet ja pilkut vaihdetaan japanilaisiksi versioikseen. Käytetyt erikoismerkit perustuvat japanilaiseen näppäimistöasetteluun, mutta ei ole mitään syytä miksei ohjelma voisi tarjota mahdollisuutta vaihtaa ne muille asetteluille helpommiksi. Välilyönnit eivät ole pakollisia, mutta niitä voidaan käyttää joissain tapauksissa sanojen rajojen merkkinä. Jikara-syntaksissa käytetyt erityismerkit esitetään taulukossa 1.

| merkki | toiminto |
|---|---|
| [ ] | katakana-tavumerkistö |
| @ | kanji-merkki |
| \ | karkausmerkki |
| <> | latinalainen-merkistö |

Taulukko 1. Jikara-erityismerkit.

## 4.3 Tavumerkit

Kaikki teksti jota ei ole muuten merkitty tai rajattu oletetaan hiragana-tavumerkeiksi.

chotto = ちょっと

Teksti hakasulkeiden sisällä oletetaan katakana-tavumerkeiksi. Toistetut vokaalit muutetaan vokaalinpidennysmerkeiksi, jollei niitä erota välilyönti tai muu merkki.

[shiiru] = シール

## 4.4 Kanjit

Kanjit kirjoitetaan käyttämällä sen japanilaista ja kiinalaista lukutapaa, jos sillä on molemmat. Homofonit ovat yleisiä, mutta todella harvalla kanjilla on sama yleinen luenta japanilaisella ja kiinalaisella lukutavalla. Kanjit jotka jakavat kaikki luennat voidaan erottaa toistaan antamalla niille erityinen nimi makrolla, jota käyttämällä tekstin voi translitteroida ASCII-merkistöstä japaniksi ja toisinpäin ilman, että ohjelma sekoittaisi merkin toiseen sen luennan perusteella. Merkin voi yrittää myös kirjoittaa käyttämällä vain yhtä luentaa, mutta sen voi olettaa vain toimivan jos merkillä on vain yksi luenta tai ei ole muita merkkejä, joilla olisi sama luenta. Tämä ei sisällä kanjin jälkeisiä tavumerkkejä, vain kanjin oman luennan. Kanji-sanamerkki alkaa @ symbolilla, jonka jälkeen tulee merkin japanilainen luenta. Jos japanilaista halutaan käyttää vain japanilaista luentaa, tulee tämän jälkeen asettaa välilyönti tai tyhjät hakasulkeet.

On myös mahdollista jättää japanilainen luenta välistä ja siirtyä suoraan kiinalaiseen luentaan, jos merkillä ei ole japanilaista luentaa tai sitä ei haluta käyttää. Nämä kirjoitetaan hakasulkeisiin, koska japanilaisissa sanakirjoissa kiinalainen luenta kirjoitetaan

yleensä katakana-tavumerkeillä, jotka kirjoitetaan myös hakasulkeisiin.

$$@\text{no[in]mu} = \overset{の}{飲}む$$

$$@[\text{eki}] = \overset{えき}{駅}$$

$$@\text{watashi} = \overset{わたし}{私}$$

Käytettävät kanjit haetaan ”sanakirjatiedostosta”. Kanjeja lisätään sanakirjaan tai suoraan tekstiin makrona seuraavasti:

```
#kanji "日" "ひ|び|か" "ニチ|ジツ"
```

Ensimmäisenä tulee ilmetä makroilmaisu ”#kanji”, jonka jälkeen tulee kolme paria lainausmerkkejä. Samalla rivillä myöhemmin olevia merkkejä ei oteta huomioon. Välilyönnit eivät ole pakollisia lainausmerkkien ja kanji-direktiivin välillä. Ensimmäisen lainausmerkkiparin väliin tulee kanjimerkki. Toiseen lainausmerkkiparin väliin tulee merkin japanilainen lukutapa. Jos merkillä on useita luentoja, tulee lukutapojen väliin putkimerkki ”. Viimeisen lainausmerkkiparin väliin tulee merkin kiinalainen lukutapa, käyttäen putkimerkkiä kuten japanilaisessa lukutavassa.

## 4.5 Erikoismerkit

Jikara-translitteraatiossa piste, pilkku, huutomerkki, kysymysmerkki ja lainausmerkit muutetaan vastaaviksi japanilaisiksi erikoismerkeiksi alla olevan taulukon 2 mukaisesti. Loput merkit vaativat makrojen käyttöä, josta annetaan esimerkki myös alla. Huomionarvoista on, että japanilaiset lauseen päättävät merkit sisältävät itsessään tyhjää tilaa seuraavan lauseen aloittavaa merkkiä varten, joten välilyöntiä lauseiden välillä ei tarvita.

```
#symbol "nakaguro" "・"
```

| Jikara-merkki | Japanilainen merkki |
|---|---|
| . (piste) | 。 |
| , (pilkku) | 、 |
| ! (huutomerkki) | ！ |
| ? (kysymysmerkki) | ？ |
| ’ (heittomerkit, pari) | 「」 |
| ”(lainausmerkit, pari) | 『』 |

Taulukko 2. Erityismerkkien translitterointi järjestelmien välillä.

## 4.6 ASCII-merkit ja paenta

ASCII-merkit kirjoietaan kulmasulkeisiin. Mitään tekstiä kulmasulkeiden välissä ei translitteroida tai käsitellä.

私(わたし)は DVD を買(か)いました。

@watashi ha <DVD> wo @ka[bai]imashita.

Vaihtoehtoisesti yhden merkin voi paeta kenoviivalla. Kenoviiva tuotetaan käyttämällä kahta kenoviivaa peräkkäin.

## 4.7 Käytännön hyöty verrattuna muihin järjestelmiin

Jikara-syntaksin etu verrattuna muihin järjestelmiin on sen kyky palauttaa alkuperäinen syöteteksti japaninkielisestä tekstistä tai muuttaa alunperin japaniksi kirjoitettu teksti ASCII-enkoodatuksi tekstiksi. Vastaavia järjestelmiä on kehitetty, mutta niiden toimintatapa on erilainen tässä tutkielmassa esitetystä (Bilac ja Tanaka 2004, Goto ym. 2004, Mammadzada 2021). Vaikka syntaksi on ihmisluettavissa, sen rakenne ei tee siitä järkevää esitysmuotoa ihmisten luettavaksi. Toinen ominaisuus minkä tämä translitterointi mahdollistaa on tekstin pakkaaminen ilmän häviötä tietyissä olosuhteissa.

# 5 Teoreettinen pakkaussuhde

Tekstin enkoodauksessa Unicoden UTF-8 formaatti on hyvin yleinen. Tässä enkoodauksessa yksi merkki voi koostua yhdestä neljään tavua (Allen ym. 2012). Koska formaatti on yhteensopiva ASCII-enkoodauksen kanssa täyttävät ASCII-merkit kaikki yhden tavun mittaisten merkkien paikat. Japanin kielen tavumerkit, erityismerkit ja yleiset kanjit sen sijaan koostuvat valtaosin kolmesta tavusta ja harvinaisemmat kanjit neljästä tavusta. Tästä johtuen järjestelmä, joka pystyy translitteroimaan ASCII-syötteen ja japanin kielen UTF-8 enkoodauksen välillä pystyy mahdollisesti pakkaamaan tietoa, mutta tähän vaikuttaa vahvasti erilaisten merkkien käyttötiheys yleisessä japanin kielessä, joka on esitetty taulukossa 3.

| Merkki | Esiintymistiheys |
|---|---|
| Kanji | 41,38% |
| Hiragana | 36,62% |
| Katakana | 6,38% |
| Erityismerkki | 13,09% |
| Arabialainen numeraali | 2,07% |
| Latinalainen aakkonen | 0,46% |

Taulukko 3. Merkistöjen esiintyminen japanin kielessä (Chikamatsu ym. 2000).

## 5.1 Tavumerkit, latinalaiset merkit ja arabialaiset numerot

Koska Jikara-translitterointi tehokkaampi esitystapa hiragana-tavumerkeille, jos joukossa on alle kolmella merkillä ilmaistavia tavuja syntyy tästä tilasäästöä verrattuna UTF-8 enkoodaukseen. Katakana-tavumerkkien, latinalaisten aakkosten ja arabialaisten numeraalien kirjoittamiseen kuluu kaksi ylimääräistä tavua hiragana-tavuista rajaavien haka- ja kulmasulkeiden tallentamiseen. Katakana-tavumerkkien kohdalla, jos teksti on tarpeeksi pitkä ja sisältää alle kolmella merkillä ilmaistavia tavuja, on mahdollista, että Jikara-translitterointi on tehokkaampi ilmaisutapa. Latinalaisten aakkosten ja arabialaisten numeraalien kohdalla hävitään varmasti kaksi tavua UTF-8 ilmaisutapaan verrattuna. Erikoismerkkien kohdalla piste ja lainausmerkit pystytään automaattisesti translitteroimaan latinalaisista merkeistä japanilaisiin merkkeihin taulukon

2 mukaisesti säästäen kaksi tavua jokaista merkkiä kohden.

## 5.2 Kanjit

Kanjien osalta pakkaussuhteen määrittäminen on monimutkaista, mutta Jikara-translitterointi on aina laskennallisesti tappiollista verrattuna vastaavaan UTF-8 esitystapaan. Jikara-translitterointi syntyy yhdestä kolmeen tavua erotinmerkkejä ja jokaisesta luennan latinalaisesta merkistä yksi tavu. Tällöin Jikara-translitteroinnin esitystapa on suurempi muistissa kuin UTF-8. Jikara-translitteroinnin pakkaussuhde siis riippuu käytettävien merkkien esiintyvyydestä ja näiden suhteellisesta pakkaussuhteesta. Alla on esitetty taulukossa 4 allaolevasta enkoodausesimerkistä tavukoosta muistissa japaniksi UTF-8 enkoodauksella ja Jikara-translitteroinnilla.

私(わたし)がよくコーヒーを飲(の)みます。

@watashigayoku[koohii]@no[in]mimasu.

| Merkkijärjestelmä | Koko tavuissa |
|---|---|
| UTF-8 | 42 |
| Jikara | 38 |

Taulukko 4. UTF-8 enkoodauksen ja Jikara-translitteroinnin tavukoot ylläolevalle esimerkille.

Kuten yllä on näytetty, on mahdollista käyttää Jikara-translitterointia japaninkielisen tekstin pakkaamiseen häviättömästi. Seuraavaksi esitämme karkean laskennallisen pakkaussuhteen arvion perustuen merkistöjen suhteelliseen esiintymiseen.

## 5.3 Teoreettinen pakkaussuhde

Pakkaussuhteen toimivuutta voidaan arvioita laskemalla Japanin valtion yleisesti käytössä olevien Joyo-kanjien listauksesta (文化庁 2010) keskimääräisen kanjin luennan ASCII-enkoodattu pituus, joka on esitetty alla taulukossa 5. Tässä otettiin huomioon vain kunkin kanjin ensimmäinen merkitty luenta kuten merkitty Joyo-kanjien listauk-

sessa, jonka japanin opetus-, kulttuuri, -urheilu, -tiede ja teknologiaministeriö on julkaissut (文化庁 2010).

| Luenta | Keskimääräinen luennan pituus tavuissa latinalaisilla aakkosilla |
|---|---|
| Japanilainen luenta | 3,02433318 |
| Kiinalainen luenta | 2,33083762 |

Taulukko 5. Joyo-kanjien latinasoitujen luentojen pituus tavuissa.

Tästä voidaan laskea keskimääräisen Joyo-kanjin pituus Jikara-translitteroinnilla, joka on lisäämällä kolme tavua erotinmerkkejä 8,3551708 tavua. Lisäksi voidaan laskea tavumerkkien keskimääräinen tavukoko Jikara-translitteroinnilla olettaen, että kaikkien tavumerkkien esiintyvyys on sama. Tässä otetaan huomioon tavumerkit, niiden handakuten ja dakuten diakriittisten merkkien versiot ja digrafit. Huomiotta jätetään toistuvat konsonantit geminaatan avulla ja katakana-tavumerkkien toistuvan vokaalin merkki sekä iteraatiomerkit kaikille merkistöille. Erikokoisten digrafien ja tavumerkkien määrä on esitetty alla taulukossa 6.

| Jikara tavumerkit | Lukumäärä | Suhteellinen esiintyvyys | UTF-8 koko |
|---|---|---|---|
| 3-tavuinen digrafi | 36 | 33,64485981% | 6 |
| 3-tavuinen tavumerkki | 2 | 1,869158879% | 3 |
| 2-tavuinen tavumerkki | 63 | 58,87850467% | 3 |
| 1-tavuinen tavumerkki | 6 | 5,607476636% | 3 |

Taulukko 6. Tavumerkkien esiintyvyys ja tavukoot.

Tästä voidaan laskea, että keskimääräinen koko tavumerkille tai digrafille on UTF-8 enkoodauksessa 4,00935679 tavua ja Jikara-translitteraatiossa 2,29906542 tavua. Yhteenveto merkkien keskimääräisistä ko'oista on esitetty alla taulukossa 7.

Yhdistämällä tämän tiedon taulukon 3 tietoihin, voidaan laskea karkea arvio Jikara-translitteroinnin pakkaussuhteelle edellämainittujen oletusten piirissä. Tässä oletetaan, että erityismerkkien määrä joita Jikara-translitterointi ei tue on mitätön verrattuna merkkeihin joita se tukee. Tällöin voidaan hypoteettiselle sadalle merkille, joiden suhteet ovat samat kuin taulukossa 3 laskea näiden merkkien koko UTF-8 enkoodauksella ja Jikara-translitteraatiolla. Tulokset teoreettiselle pakkaussuhteelle merkkityypeittäin

| Merkki | Keskimääräinen koko (Jikara) | Keskimääräinen koko (UTF-8) |
|---|---|---|
| Kanji | 8,3551708 | 3 |
| Tavumerkki/digrafi | 2,29906542 | 4,00935679 |
| Erityismerkki | 1 | 3 |
| ASCII-merkki | 1* | 1 |

Taulukko 7. Merkkien keskimääräinen koko UTF-8 enkoodauksella ja Jikara-translitteraatiolla. **Latinalaiset merkit ja arabialaiset numeraalit vaativat kaksi tavua erotinmerkkejä.*

ja yhteensä on esitetty taulukossa 8.

| Merkki | UTF-8 | Jikara | Pakkaussuhde |
|---|---|---|---|
| Kanji | 124,14 tavua | 345,736968 tavua | 278,505693% |
| Tavumerkki/digrafi | 172,402342 tavua | 98,8598131 tavua | 57,3425% |
| Erityismerkki | 39,27 tavua | 13,09 tavua | 0,333333...% |
| ASCII-merkki | 2,53 tavua | 2,53 tavua | 100% |
| Koko yhteensä | 338,342342 tavua | 460,216781 tavua | 136,021042 % |

Taulukko 8. Teoreettinen pakkaussuhde Jikara-translitteroinnilla

Kuten tuloksista voidaan todeta Jikara-translitterointi ei sovellu ainakaan tässä hypoteettisessa mallissa japanin kielen pakkaamiseen verrattuna pelkkään UTF-8 esitystapaan. Kuitenkin tässä on oletettu, että kaikki merkit esiintyvät omassa kategoriassaan yhtä suurella todennäköisyydellä ja on mahdollista, että todellinen pakkaussuhde luonnollisessa tekstissä voi poiketa tässä lasketusta teoreettisesta pakkaussuhteesta.

# 6 Yhteenveto

Tässä tutkielmassa esiteltiin japanin kielen ja kieliopin erityispiirteet koskien translitteraatiota latinalaisiin aakkostoon. Tunnistamalla kielen ja kieliopin taustalla olevan logiikan tutkielmassa esitettiin enkoodaus/translitteraatio-menetelmä, Jikara, joka poiketen muista järjestelmistä on reversiibeli ja deterministinen. Jikara-translitteraation heikkouksia on sen lisäämä näppäilyn tarve ja kömpelömpi tapa esittää kanji-merkkejä, koska se vaatii yleensä kahden luennan kirjoittamisen. Tätä järjestelmä pyrkii kompensoimaan mahdollistamalla käyttäjäkohtaisten makrojen käytön, joita voidaan käyttää myös erityismerkkien ilmaisuun.

Lisäksi tutkielmassa tutkittiin lyhyesti esitellyn translitteraatiojärjestelmän käyttöä tekstin pakkaamiseen, mutta käytetyllä laskutavalla translitteroidun tekstin koko muistissa on suurempi kuin vastaava japaninkielinen UTF-8 enkoodattu esitystapa. Tästä huolimatta käytetty laskutapa oli hyvin yleistävä ja ei ottanut huomioon erilaisten merkkien esiintyvyyttä omassa merkistössään ja näiden vaikutusta pakkaussuhteeseen.

# Lähteet


Allen, Julie D, Deborah Anderson, Joe Becker, Richard Cook, Mark Davis, Peter Edberg, Michael Everson, Asmus Freytag, Laurentiu Iancu, Richard Ishida ym. 2012. "The unicode standard". *Mountain view, CA,* 660–664.

Bilac, Slaven, ja Hozumi Tanaka. 2004. "A hybrid back-transliteration system for Japanese". Teoksessa *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics,* 597–603.

Chikamatsu, Nobuko, Shoichi Yokoyama, Hironari Nozaki, Eric Long ja Sachio Fukuda. 2000. "A Japanese logographic character frequency list for cognitive science research". *Behavior Research Methods, Instruments, & Computers* 32 (3): 482–500.

Eells, Walter Crosby. 1952. "Language reform in Japan". *The Modern Language Journal* 36 (5): 210–213.

Frellesvig, Bjarke. 2010. *A history of the Japanese language.* Cambridge University Press.

Goto, Isao, Naoto Kato, Terumasa Ehara ja Hideki Tanaka. 2004. "Back transliteration from Japanese to English using target English context". Teoksessa *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics,* 827–833.

Henshall, Kenneth G, Christopher Seeley ja Hank De Groot. 2003. *Guide to Reading & Writing Japanese.* Tuttle Publishing.

Itkonen, Terho, ja Sari Maamies. 2002. *Uusi kieliopas: Terho Itkonen; tarkistanut ja uudistanut Sari Maamies.* Tammi.

Mammadzada, Sabina. 2021. "A review of existing transliteration approaches and methods". *International Journal of Multilingualism,* 1–15.

Yamaguchi, Toshiko. 2007. *Japanese language in use: An introduction.* A&C Black.

文化庁. 2010. 常用漢字表. 文部科学省.

日本産業標準調査会. 2000. “仮名漢字変換システムのための英字キー入力から仮名への変換方式”. *JISX4063*.