

クリンゴン語翻訳システムを目指して — 解析編 — *

電子技術総合研究所

高橋直人

ntakahas@etl.go.jp

1 はじめに

機械翻訳を含め一般の自然言語処理システムは、通常、形態素解析 → 統語解析 → 意味解析 という順に処理が進むため、形態素的・統語的な処理が終了するまでは意味処理に着手することができない。ところが自然言語には、大規模語彙・多彩な文法規則・表層表現とは異なる意味を持った慣用句の存在、等の特徴があり、そのため与えられた自然言語に属するいかなる文をも形態素・統語解析することが可能であるような計算機プログラムを作成することは事実上不可能である。

このため自然言語処理研究においては、ある自然言語の部分集合を研究対象とすることが多い。つまり、形態素解析・統語解析が可能な文のみを処理対象とするのである。しかし、入力文に制限を加えることは、同時にある種の意味構造を考慮の対象から排除してしまう危険性も持っている。

一方、たとえ文法構造等に人為的な制限を加えた人工言語であろうと、その言語の完全翻訳システムをつくることは、一般の機械翻訳に必要な意味構造およびその表現方法に関する研究を大きく前進させる可能性を持つ。なぜなら、エスペラント等いくつかの人工言語¹は実際に人間同士のコミュニケーションの道具として十全に機能することが経験的に実証されており、それゆえこれらの人工言語における意味表現能力は、一般の(いわゆる)自然言語に劣るものでないと言えるからである。

クリンゴン語は「架空の自然言語」という設定であるので、エスペラント等の人工言語とは異なり、語彙的・統語的曖昧性が(おそらく意図的に)残されている。そういった意味で自然言語を単純化した一種のモデルと考えられる。また語彙、統語規則、用例が極めて限定されている — 悪くいえば貧弱である — ため、日本語その他の自然言語からクリンゴン語への翻訳は人間にとってもしばしば困難であるが、反対にクリンゴン語から自然

言語への翻訳は努力次第で機械化が可能であると考えられる。

以上の考えのもとに、著者はクリンゴン語翻訳システムを構築中である。すでに形態素・統語解析プログラムのパイロット版は完成し、現在はそれによって得られた解析結果を機械翻訳用中間言語へ変換するための規則を整備している段階である。

本稿ではまずクリンゴン語について簡単に説明し、次いで形態素・統語解析プログラムの解説を行う。さらに中間言語表現とそれをクリンゴン語に適用した場合の問題点について述べ、最後に今後の予定について記述する。

2 クリンゴン語

クリンゴン語とは、米パラマウント社の人気TV・映画シリーズである「スタートレック」に出てくる異星人の言語である。当然架空の言語ではあるが、言語学者によって基本的な語彙・文法が設定されており、書籍として出版されている [1]。また学習用カセット [2][3]、ことわざ集 [4] 等も入手可能である。さらに Klingon Language Institute という団体 [5] が、聖書やシェークスピア作品等の翻訳や通信教育を通じて普及活動を行っている。また米国以外にも熱心な学習者がおり、インターネットを通じて活発に情報発信を行っている [6][7]。

クリンゴン語はその目的からして英米語話者にとって奇妙に感じられる様に設計されている。主な特徴を以下に挙げる。

- 発音上は、声門閉鎖を含め喉音が多い(アラビア語的)。しかし地球人に発音できない音はない。
- 語は基本的に単音節から成り(中国語的)、子音-母音-子音という構造をとる。
- 名詞および動詞に各種接尾辞が後続して膠着語的性質を示す。日本語によく似た助動詞や助詞も存在する。(図 1, 2)

*Aiming at a Klingon translation system — parsing — by TAKAHASHI Naoto

¹広い意味ではインドネシア語その他も人工言語と言えよう。

QaghHommeyHeylljmo'
 おまえの小さな間違いらしきもののせいで
 Qagh [名詞] 誤り
 -Hom [語尾] 縮小辞 (小さな)
 -mey [語尾] 複数
 -Hey [語尾] 確信度 (～らしい)
 -llj [語尾] 二人称単数所有形 (おまえの)
 -mo' [語尾] 理由 (～の故に)

図 1: 名詞に接尾辞が後続する例

HeghqangmoHlu'pu'
 彼 [女] は死にたくなかった
 (それによって彼 [女] は死にたくさせられた)
 (無形) [人称接頭辞] 三人称単数が三人称単数を
 Hegh [動詞] 死ぬ
 -qang [語尾] 希望 (たい)
 -moH [語尾] 使役 (させる)
 -lu' [語尾] 受動 (られる)
 -pu' [語尾] 完了 (た)

図 2: 動詞に接尾辞が後続する例。クリンゴン語の人称接頭辞は、主語と目的語の両方の人称・数を示す。

- 基本語順は OVS(図 3)。それ以外の副詞や後置詞句は一般に文頭に置かれる。

語彙としては 2000 語弱が定義されているのみである。SF に特有の語彙が充実している一方で、通常我々が基本語と考える語の多くが未定義であるなど、アンバランスである感は否めない。

3 パーザ

形態素・統語解析用として、Prolog による bottom-up shift-reduce パーザ [8] を作成した。このパーザには文献 [1] に出てくる文法項目のうち、第 7 章 Clipped Klingon (軍隊内部で用いられる命令伝達用口語という想定) を除いたすべてが組み込まれている。したがって適切な辞書さえ用意すれば、文献 [1] で「公式の書き言葉」とされたクリンゴン語はすべて解析可能のはずである。もちろん、このパーザだけですべての語彙的・統語的曖昧性を解消することができるわけではないが、少なくとも出力の中に必ず正しい解析木が含まれることは期待できる。

puq legh yaS
 士官は子供を見る
 puq [名詞] 子供 (目的語)
 legh [動詞] 見る
 yaS [名詞] 士官 (主語)

図 3: クリンゴン語の基本語順：目的語-動詞-主語

クリンゴン語パーザを作成するに際しては、空白による語の区切りと、統語的な区切りがずれる例があることに注意する必要がある。たとえば「街の中で」という後置詞句は “vengDaq” (veng:町, -Daq:～の中で) と表されるが、これが「大きな街の中で」になると “veng tInDaq” (tIn:大きい) のように名詞接尾辞 -Daq は名詞 veng ではなくて形容詞 tIn に付加され、しかも名詞と形容詞との間にはスペースが置かれるようになる。

ことによると係り受けという概念を用いることでこの現象に適した解析方法を考案することができるかもしれないが、これについては今後の研究にゆだねる。

パーザの出力例を図 4 に示す。

4 中間言語

機械翻訳のためには、文の内部構造を解析するだけではなく、そこに表されている意味を何らかの形式で表現する必要がある。現在、クリンゴン語文を機械翻訳用中間言語に変換する規則の設定と、それを前節のパーザに組み込む作業を行っている。

中間言語としては CICC の近隣諸国間機械翻訳システムプロジェクトで用いられたもの [9, 10] を採用することとした。この中間言語は文の表す情報を、

1. 文の表す概念の情報
2. 話者の視点を表す情報
3. 話者や主体の意図、気持ちや判断を表す情報
4. 文章の構造を表す情報

の 4 つに分けて表現する方針を取っており、このうち 1. としては EDR 概念辞書と同一の枠組を採用している。

CICC の中間言語を使うことの利点としては、

- 日本語、中国語、タイ語、インドネシア語、マレーシア語、英語という系統の異なる 6 言語間の相互翻訳を目的に設計されたものであり、同系統の言語間

```
4 ?- parse(['Qu''vaD', 'lI'', 'De''vam'], [s0(X\
,_) ]), display_tree_as_outline(X).
```

```
s0
s
pp
np
n
Qu'-mission
p
vaD-for
s
vp
vpre
0-3-0
v
lI'-transmit data
np
n
De'-information
nsuf
vam-this
```

```
X = s0(s(pp(np(n('Qu\''-mission)), p(vaD-for)), \
s(vp(vpre(0-'3-0'), v('lI\''-'transmit data'))), \
np(n('De\''-information), nsuf(vam-this)))));
```

```
s0
s
pp
np
n
Qu'-mission
p
vaD-for
s
vp
vpre
0-3-0
v
lI'-be useful
np
n
De'-information
nsuf
vam-this
```

```
X = s0(s(pp(np(n('Qu\''-mission)), p(vaD-for)), \
s(vp(vpre(0-'3-0'), v('lI\''-'be useful'))), np(n\
('De\''-information), nsuf(vam-this)))));
```

```
No
5 ?-
-J:++ +prolog+ (Inferior Prolog:run)-L
Mark set
```

図 4: Qu'vaD lI' De'vam「この情報は作戦にとって有用だ」(Qu':作戦、-vaD:~のために、lI':役に立つ、De':情報、-vam:この)の解析結果。lI'と書かれる動詞には「役に立つ」と「データを送信する」の2つがあるため、解析には曖昧性が残る。

の翻訳を目的とした中間言語よりは柔軟であると考えられる。

- もしクリンゴン語からこの中間言語への変換が実現できれば、近隣諸国間機械翻訳プロジェクトによって蓄積されたノウハウを利用してクリンゴン語からこれら6言語への翻訳システムが実現されう。

等があげられる。

Qu'vaD lI' De'vam
(この情報は作戦にとって有用だ)

Qu'	作戦
-vaD	~のために
lI'	役に立つ
De'	情報
-vam	この

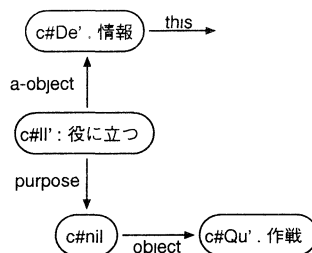


図 5: クリンゴン語から中間言語への変換例。Qu'(作戦)は有意志体でないので beneficiary ではなく purpose を使う。さらに purpose はコト概念を直接指すことはない、空概念 c#nil を補う。

一方問題点としては、

- 「転送装置」「ワープエンジン」等、スタートレックあるいはSFに特有の単語概念を表す概念見出しが存在しない
- クリンゴン語に特有の表現 — たとえば最終目的を持った進行相とそうでない進行相とでは動詞語尾が異なる — を表現するのが困難である

などが考えられるが、いずれも既存の中間言語を用いる限り避けられない問題であり、今回採用した中間言語に固有のものではない。

解決手段であるが、まず特別な語彙に関しては、独自の概念見出しを付与する以外にないのでそのようにする。また相の問題に関しては、最終目的を持った進行とそうでない進行はクリンゴン語以外の言語では区別されない、クリンゴン語から他の言語への翻訳に際してこの種の情報が落ちるのはやむを得ないと割り切り、中間言語ではどちらも単なる進行相に変換することになっている。将来、中間言語からクリンゴン語への変換を行う場合は、目的語や文脈のより詳しい分析によって適切な相を選択する必要が生じると予想される。

前述のクリンゴン語文(図4)を中間言語に(手で)変換した例を図5に示す。

5 おわりに

現在の方針としては、まず中間言語への書き換え規則を完成させ、それをパーザに統合することを第1の目標としている。

次に文献 [1] 以降断片的に発表された新しい表現、特に句読点の用法に関する文法規則を追加し、可能ならば高速化・移植性向上のために C あるいは Java でパーザを書き直したいと考えている。

パーザの一般公開も考えているが、著作権の関係 (すべての語彙が著作物である) から慎重に対処する必要があると思われる。

参考文献

- [1] Marc Okrand: *The Klingon Dictionary*, Pocket Books, New York, 1992.
- [2] Barry Levine and Mark Okrand: *Conversational Klingon*, Simon & Schuster, New York, 1992.
- [3] Barry Levine and Mark Okrand: *Power Klingon*, Simon & Schuster, New York, 1993.
- [4] Mark Okrand: *The Klingon Way*, Pocket Books, New York, 1996.
- [5] <http://www.kli.org/>
- [6] <http://www.asahi-net.or.jp/~VZ4S-KUBC/ikic.html>
- [7] <http://www.bahnhof.se/~dark/klingon/>
- [8] M. A. Covington: *Natural Language Processing for Prolog Programmers*, Prentice-Hall, New Jersey, 1994.
- [9] 財団法人 国際情報化協力センター 機械翻訳システム研究所: 中間言語表現法 第2版 6-CICC-MT37, 東京, 1995.
- [10] Machine Translation System Laboratory, Center of the International Cooperation for Computerization: *Interlingua Final Edition* 6-CICC-MT36, Tokyo, 1995.