自然言語処理　５３－８
（１９８６．１．３１）

スワヒリ語・イヌイット語翻訳システムの開発
－多言語翻訳をめざして－

西野　文人　　内田　裕士
（富士通研究所）

1. はじめに

様々な機械翻訳システムが開発されている．しかし，その多くは日本語，英語を取り扱うもので，多言語間翻訳システムを実現し，それを論じているものは数少ない．機械翻訳システムATLAS/Ⅱは多言語間翻訳を目指しており，日英システムはすでに商品化されている．そして，科学万博では英語の解析，日本語，フランス語，ドイツ語の生成システムを作成しデモをおこなった．さらに，今回スワヒリ語（解析・生成）とイヌイット語（解析・生成）の翻訳システムの試作の機会を得た．本稿ではこれらのシステムの作成経験をもとに，多言語間翻訳システム開発のあり方について論じる．

2. スワヒリ語・イヌイット語とは

スワヒリ語はケニヤやタンザニアを中心にしたアフリカ東部で話されている言語である．スワヒリ語の文章例を表１に，スワヒリ語の特徴を表２に示す．

一方，イヌイット語はカナダのフロビッシャベイなどのエスキモーによって話されている言語である．イヌイット語の文章例を表３に，イヌイット語の特徴を表４に示す．

Hakuna baridi huko? 持たない　寒い　そこ Huku ni mke wangu. これ　　妻　私の Mimi nimeoa. 私　結婚している Mimi ni direkta katika idhaa ya televisheni. 私　　ディレクタ　　テレビ局

表１　スワヒリ語　文章例

名詞は接頭辞＋語幹で構成され，活用は語頭におこる．

名詞にはいくつかのクラスがあり，接頭辞で区別される．
例）　Mtoto (child), WAtoto (children),
KItoto (childish), Utoto(childhood),
Mtu (man), WAtu (men), KItu (thing), VItu (things)
Utu (manhood), JItu (a huge man), MAJItu (huge men)

形容詞は名詞に後続して，語頭が名詞のクラスに応じて交替する．
例）　kisu kidogo kimoja (one little knife)
visu vidogo viwili (two little knife)

冠詞に相当するものはない．

動詞は［主語接頭辞］＋［時制接頭辞］＋［関係代名詞接頭辞］＋［目的語接頭辞］＋語幹＋語尾によって構成される．

名詞のクラスによって動詞，形容詞，指示詞，所有代名詞等の接頭辞が交替する．

名詞がなくても完全な文章が作られる．
例）Kitatosha (It will be enough.)
Vilitosha (They were enough.)

疑問文でも語順は変らない．
例）Viti vinatosha. (There are enough chairs.)
Viti vinatosha? (Are there enough chairs?)

語順は比較的自由である．

表２　スワヒリ語の特徴

Qiunnanngilaq silami?
寒くない　　そこ

Nuliala una.
私の妻　これ

Katititausimajunga.
私は結婚している．

aulatsijiujunga tiiviilirivvingmi.
私はディレクタだ　テレビ局で

表３　イヌイット語　文章例

語幹に派生的な接尾辞が後続して，意味や機能上の修飾・変更を加えて，最後に文法関係を示す屈折的な接尾辞がついて一つの語を形成している．すなわち一つの単語が多くの概念を組み込んだ『文』に相当している． 例）"katitititausimajunga"という単語は，動詞幹katitititauにsimaとjuとnga というう接尾辞が後続して一つの単語を形成している． 様々な接尾辞が語幹につきうる． 語順はきわめて自由である．しかし辞順（語の内部における接尾辞が配置される順序）はそれぞれの接尾辞の意味的・機能的な特性によって定まっている． 目的語に関する人称で接尾辞が変化する．

表４　イヌイット語の特徴

3. ATLAS における新言語翻訳システムの開発

翻訳システムに新しい言語の解析／生成システムを構築しようとする場合，ATLAS/Ⅱでは以下のような作業を必要とする．

① 単語辞書の開発
② 形態素連接関係規則の開発
③ 解析文法の開発
④ 生成文法の開発
⑤ 共起関係規則の開発

さらに，翻訳の質を上げるために次の作業を必要とする場合がある．

⑥ 概念構造変換規則の開発

以下，スワヒリ語，イヌイット語の解析・生成システムの開発経験を述べる．

3.1. 単語辞書の開発

単語辞書は機械翻訳システムにとって必要不可欠なものであるが，この辞書内に記述される情報の構造は翻訳プロセッサや文法に依存する．しかし，開発の初期の段階では辞書への情報の登録の仕方が確立されていないことが多い．したがって，最初に作る単語辞書は機械翻訳システムの単語辞書とは独立なものを作成できるようにすることが大切である．すなわち，機械翻訳システム用の辞書と一般の辞書の２本立てとして，利用者からみた辞書はなるべく一般的な辞書と類似しているものとした．そして，翻訳システムで必要な約束事は一般辞書から機械翻訳用辞書へ変換する際の解釈ルーチンで吸収するようにした．そこで，単語を機械翻訳システムの単語辞書として直接作成せずに，まずは単語辞書記述用シートを作成し，形態素単位，

品詞，品詞細分類（活用の違い等）といったものを記述できるようにし，これを言語精通者に記述してもらい，作成した．この作業中で必要になった文法属性等は適宜加えていき，このシートを更新していった．

このような作業の後，この辞書をATLAS 用の辞書に変換する必要があった．このシートに記述されたものをもとに必要な文法属性，形態素属性等を決定し，形態素連接関係規則を作成し，単語辞書を作成した．

3.2. 形態素分割

多言語間機械翻訳システムを目指したとき，形態素処理をどうするかは大きな問題である．ATLAS/Ⅱでは各言語ごとの形態素処理ルーチンを設けず，１つのメカニズムで，さらに，解析辞書と生成辞書の区別をせずに，様々な言語の形態素処理を行うことができるようにしている．そのため，ATLAS/Ⅱでは一般的には単語を活用無変化部と変化部に分けてそれぞれを形態素単位とし，形態素間の接続可能性は形態素連接関係規則に記述することにしている．例えば，日本語では『読む』は『読』を形態素単位として，それに『む』が接続するものとしている．英語では，"study" に対しては"stud"を形態素単位として，これに"y" や"ies" が接続するものとして活用を扱っている．このように辞書の形態素単位を作成し，あとは形態素連接関係規則を作成することにより，形態素処理を行っている．

スワヒリ語では，mtoto(child)，watoto(children)，kitoto(childish)，utoto(childhood)とか，mtu(man)，watu(men)，kitu(thing)，vitu(things)，utu(manhood)のように語幹に接頭辞がついて名詞を形成する．したがって，-toto，-tuを形態素単位として登録し，これらにm-,wa-,ki-,u- が接頭辞としてつくとして登録することにした．ただ，このときmtoto とwatotoはCHILD という概念を表わすtotoに単数接辞m または複数接辞waがついたものとみなすことで１単語にまとめることができるが，mtoto とkitotoは違う概念を表わすものなので，別々のクラスの異なる単語として登録した．同じクラスでも語幹が母音で始まるような場合には異なる接頭辞がつくことがあるが，これも連接情報を変えることで対処することにした．（chumba(ROOM)はkitu(THING) と同じクラスに分類される単語なので同じ文法属性を持つが，形態素単位umbaにはchがtuにはkiがと異なる接頭辞がつくので，異なる連接情報を持たせる．なお，この連接情報の違いは解析文法や生成文法では意識されない．）さらに，接頭辞によって語頭が変化するような単語は変化する語頭を取り除いたものを語幹として，取り除いたものまでを含めたものまでを接頭辞とした．（形容詞-ingine(OTHER)はMAクラスの複数の単語を修飾するとmengine となるので，語幹はngine とする．）

イヌイット語は１語が非常に長く，我々が普通，文として表現するような概念が１語で表現されている．（例えば，『私は結婚している．』というような文章が"katititausimajunga"といった1 語で表現されてしまうのである．）このような言語では，すべての単語を登録することは当然不可能である．１単語を細かく分割し，形態素単位を取り出す必要があった．一般に

イヌイット語では単語は，語幹＋派生接尾辞＊ ＋屈折接尾辞　の形をしているので，例えば，"katititausimajunga"は，katititau （結婚する），sima（残存結果を示す），ju，nga （私）というように分割することができる．これらをそれぞれ形態素単位として登録するが，さらに，前後の音素で文字が交替する接尾辞に対しては，それを無変化部と変化部に分割したものをそれぞれ形態素単位とした．接尾辞の辞順は意味的・機能的に決っているので，形態素連接関係規則で接尾辞どうしの連接可能性を定義し，正しい形態素分割，形態素生成ができるようにした．

3.3. 文法開発

3.3.1. 解析文法の開発

ATLAS/Ⅱでは，まず入力テキストを形態素連接関係規則をもとに接続可能な形態素単位に分割する．その後の形態素解析や構文解析は解析文法が行う．ATLAS/Ⅱの解析システム（esper ）は隣り合う２つの単位の文法属性の集合の組合せから次のアクションを決定していくものである．

スワヒリ語解析文法では，解析フェーズを大きく２つに分割し，形態素解析（空白で区切られた単語内の処理をする）を実行した後に構文解析を実行するようにした．形態素解析では，接頭辞＋語幹をまとめあげ，空白等の構文解析に不要なものを除去する．形態素解析が終了した時点では，単語の列になっている．スワヒリ語では動詞に主語接頭辞や目的語接頭辞がつくので，主語や目的語の語順は比較的自由である．したがって，構文解析ではこの情報をもとに主語や目的語を決定する必要があった．さらに，主語や目的語が省略されている場合には動詞の接頭辞の情報からそれらを補う必要もあった．

イヌイット語では１つの単語が多くの概念を含む接辞で構成されているので，上記のような解析方法をとらず，形態素解析と構文解析を同時に行った．イヌイット語でも，接尾辞の変化から主語や目的語を決定する必要があった．

3.3.2. 生成文法の開発

ATLAS/Ⅱでは構文生成と形態素生成を区別していない．したがって，空白を出力したり，大文字化したり，語尾変化をさせたりするのはすべて生成文法でおこなう．これは大きな負担のようにみえるが，生成文法がモジュール化できるので，たいして問題にはならない．特に今回のイヌイット語のような言語では，この構文生成と形態素生成を区別していないことが役立った．

スワヒリ語では名詞のクラスに応じて形容詞や動詞の接頭辞が交替するので，名詞のクラスや単複の情報は，その名詞を修飾している概念の単語にその旨を知らせる．修飾している形容

詞はその情報により，接頭辞を選ぶ．

例　小さなナイフ

小さい　　語幹dogo

ナイフ　　語幹su

この文ではナイフは複数ではないので単数の接頭辞を出力するように要求すると，システムは各クラスの名詞単数接頭辞をもってくる．ここで，このsuの隣接番号，接頭辞の隣接番号と形態素連接関係規則から，KITUクラスの単数接頭辞kiのみが接続可能になり，これを出力する．（その際，KITUクラスを出力したという情報は受け取っている．）その後自分自身の形態素単位suを出力し，さらに空白を出力する．その後生成が進み，このナイフを修飾している概念に出力を要求する時には，単数という情報とKITUクラスであるという情報を送る．これにより形容詞語幹dogoにはKITUクラス単数接頭辞kiを前につけて出力する．図５にATLAS/Ⅱのスワヒリ語名詞の形態素生成の文法を示す．

```
複数形を要求                          ／<pl>／アウトアーク処理／＊；
接頭辞を要求，￢複数形を要求／＊／出力／名詞単数形接頭辞；
接頭辞を要求，複数形を要求　／＊／出力／名詞複数形接頭辞；
＊                                    ／＊／自分自身出力／＊；
￢文の終了                            ／＊／出力／空白；
```

図５　ATLAS/Ⅱのスワヒリ語名詞形態素生成文法

3.3.3. 文法開発の進め方

文法には，解析文法と生成文法を全く同じものとしているシステムもあるが，ATLAS/Ⅱをはじめとする多くのシステムでは，解析文法と生成文法は異なるものになっている．今回の試作では，文法開発は，解析文法と生成文法を同時に作成した．しかし，生成文法の方をまず重点的に開発することにした．ATLAS/Ⅱにはすでに日本語解析をする能力があったので，日本語→スワヒリ語，日本語→イヌイット語というように日本語解析の結果を利用することによって，生成文法を作成することができた．文法の記述は，まず言語精通者が翻訳システムとは独立に文法を記述する．これを技術者が文法記述言語で記述する．例文を翻訳システムにかけて翻訳し，その結果を見て誤りがある場合には生成過程を言語精通者に説明し，どのような条件が不足しているかを調べ文法を修正する．このようにして文法を作成していった．

生成文法に重点を置いて先に開発したのは次のような理由からである．

① 解析文法の難しさは，その言語特有の現象を扱うところよりも，自然言語一般に共通する問題を取り扱うところにある．これに対して，生成文法はその言語の現象をいかに翻訳システムの文法記述言語で表現するかということにある．したがって，文法属性等

の単語辞書の情報を整備，細分化していくには，解析文法より生成文法の方に重点を置いて整備したほうがしやすいと考えた．

② 文法開発する者がその言語に精通していない者の場合，テストとして，解析用の文を自由に与えることができない．これに対して，生成は自由にいろいろな文章を試すことができる．その結果を専門家にみてもらい，誤った結果に対してはその原因となる条件や不足している情報等を聞きだし易い．

4. マイナーな言語の機械翻訳システム

現在，各社で開発が進められている翻訳システムは商用システムあるいは自然言語処理システムの研究ということで，日本語，英語といったメジャーな言語の翻訳システムが中心である．あとは，ドイツ語，フランス語，スペイン語，中国語，韓国語といったような，比較的馴染みのある言語が殆どであり，スワヒリ語，イヌイット語といった言語の翻訳システムは，ほとんど開発されていない．

これらの言語の翻訳システムを開発することに意義はあるのだろうか．まず，これらの言語の翻訳家の人は数少ない．そして，これらの言語に対する辞書，文法書なども数少なく，なかなか手にいれることができない．したがって，もしこのような言語から，あるいはこのような言語からの翻訳が必要になったときには非常に苦労することになる．またその言語に慣れていない場合，単語を辞書で探し出すことも結構大変である（特に活用がある言語では辞書に載っている原形が何かわからないことも多いし，また今回のイヌイット語のように接辞がついて語が形成されるような言語では，どこが接辞の切れ目かも知らなければならない．）．したがって，辞書引きをして，多少稚拙ながらも翻訳をするならば，かなり助けになるであろう．

限られた話し手の間でしか使われない遠からず消滅するかにみえる小言語も少なくない．わずかな話し手しかいない言語だけで，それを使う人が文明世界にひとり立ちしていけるものではない．しかし，様々な言語の機械翻訳システムが開発され，その言語を使う人に情報を流すことができれば，この危機ものり越えることができるであろう．

5. まとめ

ATLAS/Ⅱではいろいろ不都合な点もあったが，しかし，とにかくはスワヒリ語，イヌイット語の解析・生成をするシステムを作成することができた．もちろんここで作成したシステムは規模の小さなものではあるが，これを大量にしていく実験はすでに日本語や英語のシステムで経験ずみである．機械翻訳システムの能力を調べるためにはこのような特殊な言語のシステムを作成してみることも必要であろう．機能としてシステムが構成できるならば，あとは記述力

を高めるとか，高速化するといった仕事をすることになる．

今後このような各言語の機械翻訳システムを開発するには，どのように文法を開発するか，辞書をどのように管理するか（特に中間言語を採用しているシステム），目標言語側に対応する概念がない時はどうするか，外来語をどう取り扱うか，といったことを検討する必要があるであろう．

謝辞

スワヒリ語，イヌイット語の翻訳システム作成に御協力くださった東京外国語大学の守野教授，宮岡教授，早津助手，大阪外国語大学の中島助教授，及びNHK の深瀬氏，畠山氏に深謝いたします．

参考文献

(1) 宮岡伯人：エスキモーの言語と文化　弘文堂(1958)

(2) D.V.Perrott:"SWAHILI" Hodder and Stoughton Ltd.(1951)