タイトル	階層的なアイヌ語・日本語対訳データの構成と層指定解析ツールの開発
著者	安曇，恭徳；桃内，佳雄
引用	北海学園大学工学部研究報告，36：175-193
発行日	2009-02-20

階層的なアイヌ語・日本語対訳データの構成と層指定解析ツールの開発

安 曇 恭 徳*・桃 内 佳 雄*

Development of Analysis Tool for Hierarchical Ainu–Japanese Translation Data

Yasunori AZUMI* and Yoshio MOMOUCHI*

あらまし

アイヌ語・日本語機械翻訳システムの開発やアイヌ語と日本語の対照言語学的な考察にとって，アイヌ語・日本語対訳データの構成とその利用のしくみの構築は有用である．本報告では，アイヌ語・日本語対訳データの階層的な構成と問題点について述べた後，階層的な構成を考慮に入れたデータ解析のための層指定解析ツールの開発と応用について報告する．層指定解析ツールの開発は，オブジェクト指向に基づくモデリング言語UMLを用いて設計を進め，オブジェクト指向型言語Javaにより実現しており，そのモデリングの概要についても述べる．

1．はじめに

アイヌ語・日本語機械翻訳システムの開発やアイヌ語と日本語の対照言語学的な考察にとって，アイヌ語・日本語対訳データの構成とその利用のしくみの構築は有用である．本報告では，アイヌ語・日本語対訳データの階層的な構成と問題点について検討した後，階層的な構成を考慮に入れたデータ解析のための層指定解析ツールの開発と応用について報告する．具体的な応用の例として，単語の多義性の解消と複合語の推測の解析例について述べ，その有用性について検討する．また，層指定解析ツールは，オブジェクト指向に基づくモデリング言語UML（Unified Modeling Language）を用いて設計を進め，オブジェクト指向型言語Javaによって実現している．層指定解析ツールの構成をUMLで定義されているモデリングのための基本的な図式であるクラス図とシーケンス図を用いて表現することの有効性についても確認する．

* 北海学園大学大学院工学研究科電子情報工学専攻
* Division of Electronics and Information Engineering, Graduate School of Engineering, Hokkai–Gakuen University

2．階層的なアイヌ語・日本語対訳データの構成

2.1 階層的な対訳データの構成

アイヌ語・日本語対訳データは，アイヌ語・日本語対訳データ要素の順序付けられた集合として構成される．対訳データ要素の順序は，対訳データ構成のもとになる原データの構成に依存している．対訳データ要素の基本的な構成について，まず，文献[1,2]を原データとする対訳データを構成する中で検討を進めた．その結果，アイヌ語・日本語対訳データ要素の基本的な構成は，付加コードを付与したアイヌ語文，アイヌ語品詞列，日本語逐語訳，日本語品詞列，日本語文の５層の情報から構成される単位を基本構成要素として，次のように設定することとした．文献[1,2]は，アイヌ語の入門的なテキストであり，いくつかの章，節，文などから構成されており，付加コードは，そのような構成を考慮したものとなっている．

＜５層の構成＞

［第１層］付加コード01：アイヌ語文

［第２層］付加コード02：アイヌ語品詞列

［第３層］付加コード03：日本語逐語訳

［第４層］付加コード04：日本語品詞列

［第５層］付加コード05：日本語文（自然な）

文献[1,2]では，アイヌ語品詞列，日本語逐語訳を作成するためのデータとして，単語辞書，単語索引などが準備されていて利用可能である．日本語品詞列，日本語文（自然な）については，新たに作成することになる．アイヌ語品詞列と日本語品詞列の二つを含めたのは，アイヌ語と日本語の品詞の対応について一致しない部分があるということによる．その違いは対照言語学的な考察にとって有用な情報となると考えられる．文献[1,2]に対応する対訳データ要素の構成では，基本的な構成に１層を加えて，次のような６層の情報による構成が設定されることになるであろう．日本語文（原著の）は原データに含まれている日本語訳に対応している．

＜６層の構成＞

［第１層］付加コード01：アイヌ語文

［第２層］付加コード02：アイヌ語品詞列

［第３層］付加コード03：日本語逐語訳

［第４層］付加コード04：日本語品詞列

［第５層］付加コード05：日本語文（自然な）

［第６層］付加コード06：日本語文（原著の）

６層の構成の対訳データ要素の例を次に示す．

＜6層の構成の例[1)]＞

exp0100521：sonno　keraan　.

exp0100502：副詞　自動　.

exp0100503：本当に　おいしい　。

exp0100504：副詞　形容　。

exp0100505：本当においしい　。

exp0100506：とってもおいしいよ。

＜6層の構成の例[2)]＞

upa0100101：ku=pon　hi　ta　ramma　ku=siyeye　.

upa0100102：人接＝自動　形名　格助　副詞　人接＝自動　.

upa0100103：私＝小さい　とき　に　よく　私＝病気になる　。

upa0100104：接代＝形容　形名　格助　副詞　接代＝連語・自動　。

upa0100105：私が小さいときによく病気になった。

upa0100106：私が小さかった時，よく病気になりました。

第5層の日本語文（自然な）は，日本語逐語訳に近く，それからのなるべく自然な変換により得られる日本語訳であり，日本語文（原著の）は原著に含まれる原データとしての日本語訳である．日本語文（原著の）は，原著者により作成されたものであり，日本語逐語訳から少し遠い訳，かなり遠い訳，意訳などが含まれる．例えば，次のような例がある．

＜6層の構成の例[1)]＞

exp0400621：apunno　oka　yan　.

exp0400602：副詞　自動　終助　.

exp0400603：無事に　いる　なさい　。

exp0400604：副詞　自動　自動　。

exp0400605：無事にいなさい。

exp0400606：バイバイ。

＜6層の構成の例[2)]＞

upa0201101：utari　arki　kor　ku=kor　huci　kaeka　kor　uenewsar .

upa0201102：名詞・所　自動・複　接助　人接＝他動　名詞　自動　接助　自動　.

upa0201103：人々　来る　と　私＝持つ　おばあさん　糸をよる　ながら　語り合う　。

upa0201104：名詞　自動　接助　接代＝他動　名詞　自動　接助　自動　。

upa0201105：人々が来ると私のおばあさんが糸をよりながら語り合う。

upa0201106：お客が来ると祖母は糸よりをしながら四方山話をします。

最初の例では，「バイバイ」は日本語逐語訳からはかなり遠い．2番目の例の「お客」，「四方

山話」は，少し遠い，状況に依存した意味的な訳となっている．

次に，文献[3]に対応する対訳データ要素の構成では，次のような７層の情報による構成が設定される．

＜７層の構成＞

［第１層］付加コード01：アイヌ語文

［第２層］付加コード02：アイヌ語文（切替版）

［第３層］付加コード03：アイヌ語品詞列

［第４層］付加コード04：日本語逐語訳

［第５層］付加コード05：日本語品詞列

［第６層］付加コード06：日本語文（自然な）

［第７層］付加コード07：日本語文（原著の）

６層から新たに加わったのは，第２層の「アイヌ語文（切替版）」である．もとの２層から６層は３層から７層に移動する．知里幸恵による原著アイヌ語文は，区切り単位が必ずしも品詞単位ではなく，切替版はこれを品詞単位に区切ったものである．対訳データ要素の例を次に示す．

＜７層の構成の例[3]＞

yuk0100101：” Shirokanipe ranran pishkan ， konkanipe

yuk0100102：” sirokani pe ran ran piskan ， konkani pe

yuk0100103：” 名詞 名詞 自動 自動 位名 ， 名詞 名詞

yuk0100104：「 銀 滴 降る 降る まわり ， 金 滴

yuk0100105：「 名詞 名詞 自動 自動 名詞 ， 名詞 名詞

yuk0100106：「銀の滴降る降るまわり，金の滴

yuk0100107：「銀の滴降る降るまはりに，金の滴

原著に含まれているデータは，アイヌ語文，アイヌ語文（切替版），日本語文（原著の）の三つと，詳細な単語辞書（辞典）である．これらを参照しながら，３層から５層の情報を作成することになる．このように，与えられる原データの構成と内容に依存して，基本データ要素に新たな層を設定して，有用な情報を組み込みながら対訳データの構成を進めることになる．

2.2 対訳データの構成における問題点

アイヌ語・日本語対訳データの構成は，前節で述べたように，現時点では，利用可能な原データテキストを参照しながら，多くの部分を人手で進めなければならない．本節では，その過程で考えなければならないいくつかの問題点について検討する．

（1）品詞の設定について

アイヌ語の品詞については，文献[3,4,5,6,7]において，それぞれの考えに基づいて，基本的な設定が行われている．それらの文献を参照しながら，本論文では，【表1】のような品詞の設定を行う．日本語品詞における，「接代」は，「人称接辞代名詞」の略であり，アイヌ語の人称接辞に対応して新しく作ったものである．また，「判定」は「判定詞」の略である．ここで，問題は，日本語品詞における「連語」である．これは，主に，アイヌ語の複合語に対応する単語直接翻訳（逐語訳）の日本語に充てられている．アイヌ語の複合語と「連語」については，（2）で考察する．この表の設定において，アイヌ語にはあって日本語にはない品詞範疇は次の7つである．

・人称接辞　（　対応する日本語品詞：　人称接辞代名詞（接代））
・名詞化辞　（　形式名詞，名詞　）
・完全動詞　（　連語，形容詞　）
・後置副詞　（　連語　）
・連他動詞　（　他動詞　）
・連複他動詞　（　他動詞　）
・虚辞　（　（空文字）　）

また，アイヌ語の他動詞と複他動詞は，日本語では，他動詞にまとめている．連他動詞，連複他動詞は，佐藤[7]では，分離動詞と呼ばれている．これは複合翻訳単位[9]の一つとして捉えることができ，複合翻訳単位は，連他動詞以外にも存在し，複合翻訳単位の翻訳の方略についての考察が必要となる．虚辞は，アイヌ語では形があるが，日本語に対応する形がない単位である．

（2）アイヌ語単語直接翻訳が連語となる場合について

アイヌ語の単語は，複数の要素の合成により構成されていることが多く，そのような単語を複合語と呼ぶ．複合語の構成について，佐藤[7]は，名詞と動詞の「語形成」として考察を行っている．アイヌ語複合語を一つの単語として，その日本語への単語直接翻訳は，日本語としては連語となることが一般的である．ここで，連語とは，「単独でも用いられる二つ以上の単語から構成されている単位」とする．（1）での品詞対応表では，対応する日本語の品詞範疇を「連語」としている．また，複合語でないと考えられるアイヌ語単語でも，日本語に翻訳すると連語となるような場合もある．本報告の考察の基礎的なデータとしている文献[1,2]に出現している具体的な例をいくつか以下に示す．（T：田村辞典[5]，N：中川辞典[6]における日本語訳）

まず，動詞の例を示す．

apeari：自動：：火を焚く：連語・自動
T：　ape–ari　火・を焚く　　：火をたく

表1 アイヌ語・日本語品詞対応表

アイヌ語品詞	略語	例	日本語品詞略語
名詞	名詞	ape（火），huci（母）	名詞
固有名詞	固名	Sapporo（札幌）	固名
代名詞	代名	kani（私），eani（あなた）	代名
位置名詞	位名	oro（中），oka（後，うしろ）	名詞
形式名詞	形名	pe/p（もの）	形名
名詞化辞	名辞	hi（こと，の，時）	名詞，形名
疑問詞	疑問	hemanta（何），hunna（誰），hunak（どこ）	疑問
数詞	数詞	sinep（一つ），tup（二つ）	数詞
完全動詞	完動	mean（寒い），sirpirka（天気がよい）	形容，連語
自動詞	自動	ahup（入る），poro（大きい，大きくなる）	自動，形容
他動詞	他動	e（～が…を　食べる）	他動
複他動詞	複他	ere（～が…に－を　食べさせる）	他動
デアル動詞	デ動	ne（だ，である，です）	判定
連他動詞	連他	aske uk（～が…を　招待する）	他動
連複他動詞	連複	ka（si）omare（～が…に－を　加える）	他動
連体詞	連体	tan（この）	連体
副詞	副詞	sonno（本当に），apunno（無事に）	副詞
		somo（ない：否定の副詞）	助動
		iteki（な：否定の副詞）	終助
接続詞	接続	orowa（それから）	接続
助動詞	助動	rusuy（たい），a（た）	助動
		eaykap（ことができない）	連語
間投詞	間投	o（ほら），iyairaykere（ありがとう）	間投
格助詞	格助	ta（に，で），un（へ）	格助
副助詞	副助	anakne（は），ka（も）	副助
接続助詞	接助	wa（て）	接助
終助詞	終助	na（よ），ya（か）	終助
後置副詞	後副	turano（と一緒に）	連語
人称接辞	人接	ku（私），e（あなた）	接代
接辞	接辞		接頭，接尾
虚辞	虚辞	u,e	（空文字）

N：　ape ari　火を焚く

日本語の単語直接翻訳は「火を焚く」で，この品詞範疇は，「連語・自動」とする.「自動」は「自動詞」の短縮形で，連語「火を焚く」は，アイヌ語単語の品詞範疇と同じ品詞範疇にあると考える．

次に，名詞の例を示す．

cepker：名詞：：鮭皮の靴：連語・名詞

T：　cep　魚　　ker　靴（鮭の皮か鹿の皮で作ったもの，冬用）

N：　cep　魚　　ker　靴

日本語の単語直接翻訳は「鮭皮の靴」，品詞範疇は「連語・名詞」とする．文献[2]の単語索引では「サケ皮の靴」となっている．次の二つの例は，同じく，「AのB」という形の連語・名詞（名詞句）である．

sikerpe：名詞：：シコロの実：連語・名詞

sikerpeni：名詞：：シコロの木：連語・名詞

次のような 4 要素からなる複合語の例がある．

poppetaasan：自動：：汗が出る：連語・自動

T：　pop–pe–ta–asin　ボコボコ煮立つ・水分·(？)·出る　　：汗が出る，汗をかく

N：　poppetaasan　汗をかく

日本語の単語直接翻訳は「汗が出る」か「汗をかく」で，この品詞範疇は，「連語・自動」とする．

以上の例は，アイヌ語の複合語の構成要素の日本語への単語直接翻訳をそのまま合成することによりその複合語の単語直接翻訳が得られる例である．品詞範疇もアイヌ語の品詞範疇にほぼ対応している．

複合語ではないアイヌ語単語の単語直接翻訳が日本語として連語となる例を次に示す．

mim：　　T：　魚肉，魚の身（骨や皮以外の魚肉の部分）

この例の「魚の身」の品詞範疇は「連語・名詞」とする.「魚肉」は名詞とする．

eaykap：他動：：できない：連語・自動

T：[他動]...ができない，...がへただ　[助動詞的用法]　...することができない，...するのがへただ

N：[動２][助動]（～が）できない

日本語「できる」は品詞範疇は自動詞とするのが一般的である．また，この場合の「ない」は助動詞である．日本語の品詞範疇として，「連語・自動」を充てる．

連語の規定，連語の品詞範疇の設定の方法を次のようにまとめる．

【アイヌ語の単語に対する日本語訳が，日本語の単独でも用いられうる二つ以上の単語から構

成されるとき，その日本語訳を連語と考える．そして，連語の品詞範疇は，その連語が担う日本語の品詞としての機能に対応して設定する．】

3．層指定解析ツールの開発

本章では，アイヌ語・日本語対訳データの構成に基づく層指定解析ツールについて報告する．層指定解析ツールは，アイヌ語・日本語対訳データ要素中の指定した層のデータについて，単語または文字列を単位としてN–gram解析を行うツールである．機械翻訳システムの開発において，このツールを利用して頻度の高い品詞や単語の並びを発見し，その結果を多義性の解消や複合語の推測のために応用することができる．２つの応用例についても考察する．

3.1 層指定解析とは

層指定解析は，階層的な対訳データ要素について，ある特定の階層を指定して，その階層にある，単語または文字列を「単位」とする分かち書き文として構成されている文を単位に分割し，その単位の並びについてN–gram解析を行うしくみである．N–gram解析では，データ中の隣り合ったN個の単語または文字列といった単位の間の関係を「共起関係」と呼び，データ中に含まれる単位の間の「共起関係」がどの程度現れるかを集計した結果を「共起頻度」と呼んでいる．データ中に現れる共起関係のリストを「共起関係リスト」と呼ぶ．また，N–gram解析において用いられる単位のリストを「解析単位リスト」と呼ぶ．「共起関係」にあるパターンの「共起頻度」に基づいて，データの統計的な解析を行うことができる．

3.2 層指定解析ツールの機能とGUI

3.2.1 層指定解析ツールの機能

（1）層指定解析ツールの主機能

層指定解析ツールの主機能は，階層的な対訳データに含まれる対訳データ要素について，指定した階層にある単語または文字列を単位としてN–gram解析を行うことである．

（2）層指定解析ツールの補助機能

層指定解析ツールの補助機能として，次の三つの機能が組み込まれている．

① 解析単位変換機能

層指定解析で発生する解析単位をユーザがあらかじめ設定した解析単位に変換する機能である．設定する解析単位は，テキストファイルとして作成し，“対象の解析単位－変換後の解析

単位”という表形式で記述する．なお，解析処理で発生した解析単位で，設定した解析単位以外の単位は変換されずにそのまま残す．

②　ソート順序選択機能

共起関係リストのソートの順序として，共起関係の辞書引き順または共起頻度順を選択することができる．

③　結果出力表示機能

N–gram解析の結果である共起関係リスト，解析単位リストを出力表示する．また，解析した共起関係のパターン数，総数なども表示する．

3. 2. 2　層指定解析ツールのGUI

層指定解析ツールのGUI（Graphical User Interface）は次のような構成である．

図 1　層指定解析ツールのGUI

・入力ファイル名：解析する対訳データファイル．＜GUIでの例＞　Exp 6 FinalB.txt

・出力ファイル名：解析結果を出力するファイル．＜例＞　out.txt

・データの階層：対訳データ要素の階層．＜例＞　6
・解析する階層：対訳データ要素の解析する階層．＜例＞　2
・解析単位変換表ファイル名：解析単位変換用のテキストファイル．
・gram数：解析するgram数．＜例＞　2
・頻度順にソート（チェックボックス）:出力結果を頻度順にソート．＜例＞　選択
・辞書引き順にソート（チェックボックス）:出力結果を辞書引き順にソート．
・解析表示（ボタン）:解析結果を表示する．＜例＞　選択
・結果出力（ボタン）:解析結果を所定の出力ファイルに出力・保存する．
・解析結果（リストボックス）:N–gram解析の結果を選択されたソート順序で表示する．
＜例＞　階層２（アイヌ語品詞列）で２－gram解析を行い，頻度順に表示している．
対応する階層１と３での単語並びの例とともに，上位10位を下表にまとめる．

２－gram	頻度	対応する階層１と３での単語並びの例
人接－他動	62	e=ku（あなた＝飲む），ku=koyki（私＝捕る）
人接－自動	46	ku=iwanke（私＝元気です），e=arpa（あなた＝行く）
形名－デ動	41	ruwe an（そう　です），ruwe ne（の　だ）
他動－接助	36	ci=koyki wa（私たち＝捕る　て）
名詞－人接	32	usey e=ku（お湯　あなた＝飲む）
終助－.	27	na　.（よ　。），yan　.（なさい　。）
自動－接助	26	ahup wa（入る　て）
自動－形名	25	ku=popke humi（私＝暖かい　感じ）
接助－自動	24	wa sini（て　休む），kor an（て　いる）
デ動－.	20	ne　.（だ　。）

・解析単位リスト（リストボックス）：解析単位リストを表示する．
＜例＞　［デ動，人接，他動，他動・複，代名，位名，副助，副詞，助動，名詞，・・・］
この例での解析単位リストは，層指定解析ツールがデータを読み込みながら，“　”（スペース）と“＝”を区切り記号として自動的に解析単位を切り出して，解析単位変換表ファイル名が指定されていないので，変換せずにそのまま作成する．第２層はアイヌ語品詞列の並びなので，品詞と句読点等を単位として切り出している．
・処理終了表示：頻度が１以上の共起パターン数と可能な共起パターンの総数を表示する．
＜例＞　頻度１以上の２－gramパターン数164，可能な２－gramパターン数1021．

3.3 層指定解析ツールの処理

3.3.1 層指定解析ツールの処理概要

層指定解析ツールでは，入力ファイルのデータを読み込みながら，設定した階層数を区切りとして対訳データ要素ごとに分割し，対訳データを再構成する．対訳データ中の全ての対訳データ要素から指定した解析する層のデータを取り出してまとめてリストを作成する．そして，そのリストの中の各データについて，そのデータに含まれているすべての共起関係を取り出し，共起頻度を計数する．層指定解析ツールの処理概要は以下のようにまとめることができる．

① GUI画面から，入出力ファイル名やgram数，ソート方法等の情報を入力する．
② 解析表示ボタンを押すことにより解析が始まり，入力ファイルからデータを読み込む．
③ 読み込んだデータを指定した階層数で区切って，対訳データ要素を切り出す．
④ 解析単位変換表ファイル名が指定されている場合，ファイルから変換表を読み込む．
⑤ 指定されている解析層のデータを取り出してまとめ，解析層データリストとする．
⑥ 解析層データリストからデータ（分かち書き文）一つを取り出し，単位に分割する．
⑦ 分割した単位を集めて，解析単位リストにまとめる．
⑧ 分割した単位を連結して，指定されているgram数分の単位の並びを作成する．
⑨ 解析単位変換表を参照して単位の表記を変換する．
⑩ gram数分の単位の並びを共起関係として，データから共起関係を取り出す．
⑪ 取り出した共起関係を共起関係リストの中に探し，もし見つかれば，共起頻度を１加算し，見つからなければ，共起関係リストに追加する．
⑫ ⑥から⑪を繰り返し，解析層データの共起関係をすべて取り出す．
⑬ 共起関係リストをソートする．
⑭ 共起関係リストを出力する．

3.3.2 層指定解析ツールのUMLモデリング

層指定解析ツールは，オブジェクト指向に基づくモデリング言語UML（Unified Modeling Language）を用いて，そこで定義されているクラス図とシーケンス図を作成しながら設計を進め，オブジェクト指向型言語Javaによって実現している．本節では，層指定解析ツールの構成におけるクラス図とシーケンス図について説明する．クラス図は，システムの静的な側面を表す構造図であり，システムの構成要素であるクラスの仕様やクラス間の関係を記述する．ま

図 2　層指定解析ツールのクラス図

た，シーケンス図は，システムの動的な側面を表す手順図であり，クラス間の相互作用（メッセージのやりとり）を時系列で表現する．層指定解析ツールのクラス図を【図２】に示す．このクラス図に表記されているクラスの概略仕様とクラス間の関係について以下で説明する．

・層指定解析ツール・GUI画面

テキストボックスやボタンなどのGUI画面の表示を行うクラスである．GUI画面を実現するためにタイトルとボーダを持つトップレベルウィンドウのFrameクラスを継承している．また，メインメソッドもこのクラスにあり，層指定解析ツールの基点となっている．なお，クラス図ではGUIの画面に配置される変数は省略している．

・GUIデータ

gram数や入力ファイル名など，GUI画面からユーザが入力した解析に必要な情報を保持するためのクラスである．そのため，GUI画面からの情報を必要とする“対訳データ要素”や“計数”と関連を持つ．この節の説明で，クラスは“と”で括って表現する．

・計数

“層指定解析ツール・GUI画面”から，GUIデータのオブジェクトを受け取って，層指定解析処理を行わせるクラスである．対訳データ関連の処理は“対訳データ要素”以下のクラスに行わせる．また，“対訳データ要素”の処理で得られた共起関係は“共起関係リスト”に送り計数処理を行わせる．

・共起関係リスト

共起関係とその共起頻度を格納し，それらの一覧を管理するクラスである．

・対訳データ要素

対訳データに関連するすべての処理を管理するためのクラスである．そのため，“単位分割”や“対訳データ読み込み”，“対訳データ単位の管理”と関連を持ち，このクラスから他のクラスへ読み込み処理，単位分割処理，連結や変換のための単位に関する処理などを行わせる．

・単位分割

“対訳データ要素”から受け取ったデータを単位に分割するクラスである．このクラスは，分割するデータのコード部分を取り除く“コード削除”と，単位に分割するための記号をまとめた“区切り記号”のクラスを持つ．また，“区切り記号”とは単位分割の処理に不可欠なデータを持つためコンポジションの関係を持つ．

・区切り記号

単位を分割するための区切り記号を格納するクラスである．なお，“区切り記号”は区切り記号の対象文字を単位として残すかどうかを示すデータを持つ．

・コード削除

データの付加コード部分を取り除くクラスである．層指定解析ツールにおいて付加コード部分は解析処理のノイズとなるため，単位分割を処理するとき付加コード部分を削除する．

・対訳データ単位の管理

分割された単位を処理させるためのクラスである．このクラスでは，“対訳データ”から単位を受け取って，解析単位リストの作成や解析単位変換処理，単位連結処理を行わせる．そのため，“解析単位リスト”，“連結”，“表示変更”のクラスと関連を持つ．なお，単位連結処理に必要とする，gram数分の単位を取り出す処理はこのクラスで行う．

・解析単位リスト

分割された単位から，解析単位リストを作成し，リストを保持するクラスである．

・連結

“対訳データ単位の管理”で取り出された単位をつなげて一つの文字列として出力するためのクラスである．

・表示変更

“対訳データ単位の管理”で取り出された単位と，“変換表示読み込み”で読み込んだ変換前の単位の中から検索を行い，一致した変換前の単位に対応する変換後の単位を出力することで，取り出された単位を変換後の単位に置き換えるクラスである．

・ファイル読み込み

指定したファイルから１行のデータを読み込むクラスである．“対訳データ読み込み”，“変

換表読み込み”の親クラスとしてのみ存在する．

・無視キーワード

ファイルを読み込むときに発生するノイズを削除するためのクラスである．層指定解析ツールではファイルから読み取ったデータの処理を行うのでそれらの処理において不具合の原因となり得るデータをこのクラスの処理で取り除く．

・対訳データ読み込み

対訳データを読み込み，対訳データ要素を作成し保持するクラスである．対訳データを読み込むための処理や必要となるデータは“ファイル読み込み”のクラスで定義されており，このクラスはこの部分を継承している．

・変換表示読み込み

表示変換用のテキストファイルを読み込み，読み込んだデータを対象の解析単位と変換後の解析単位に分けリスト化するクラスである．表示変換用のテキストファイルを読み込むための処理や必要となるデータは“ファイル読み込み”のクラスで定義されており，このクラスはこの部分を継承している．

次に，層指定解析ツールのトップレベルのシーケンス図を【図３】に示す．ユーザがGUI画面に操作することから始まって，システムがどのように処理を進めていくかが時系列に表現されている．詳細説明は省略するが，引き続きの処理のためのシーケンス図が作成されている．

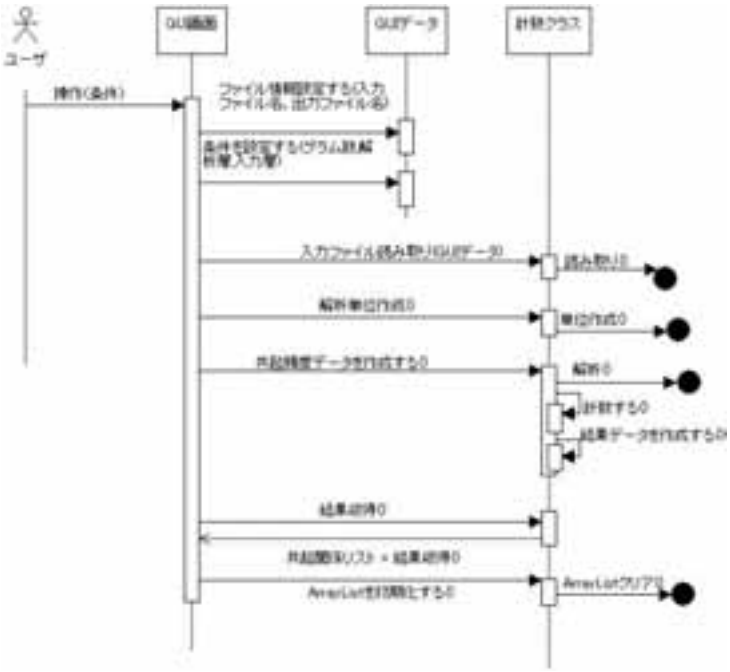

図 3　層指定解析ツールのシーケンス図（トップレベル）

以上のようなUMLモデリングにおけるクラス図とシーケンス図を基礎として，層指定解析ツールは，オブジェクト指向型プログラミング言語Javaを用いて実装されている．

3.4　層指定解析ツールの応用

層指定解析ツールを応用した具体的な解析例について検討する．

（1）多義性の解消

文献[1]に含まれる次のアイヌ語文＜1＞について，文献[1]に基づいて作成した辞書を用いて，単語直接翻訳を行うと下表のような多義を含む翻訳結果が得られる．

＜1＞　usey e=ku rusuy ya ?

アイヌ語	アイヌ語品詞	日本語	日本語品詞
usey	名詞	お湯	名詞
e	人接	あなた	接代
e	他動	食べる	他動
e	間投	はい	間投
ku	人接	私	接代
ku	他動	飲む	他動
ku	名詞	弓	名詞
rusuy	助動	たい	助動
ya	名詞	網	名詞
ya	位名	陸	名詞
ya	終助	か	終助
ya	終助	の	終助
ya	終助	ね	終助
?	区切	?	区切

上表の多義の可能性の中で，妥当と考えられる単語直接翻訳は次のようである．

usey　e　= ku　rusuy　ya　?
名詞　　人接　　他動　助動　　終助　区切
お湯　　あなた　飲む　たい　　か　　?
名詞　　接代　　他動　助動　　終助　区切

この妥当性を層指定解析ツールによる解析によって検証（推測）するために，次のような条件で層指定解析を行った．

・解析のための対訳データ：Exp 6 FinalB.txt（文献[1]に基づいて作成した対訳データ）

・解析する層：2（アイヌ語品詞列）

・gram数　：2

その解析結果は下表のようにまとめられる．

アイヌ語	アイヌ語品詞列　2－gram
usey–e	名詞－人接：32，名詞－複他：1，名詞－間投：0
e–ku	人接－他動：62，人接－名詞：0，人接－人接：0 複他－名詞：1，複他－他動：0，複他－人接：0 間投－他動：0，間投－人接：0，間投－名詞：0
ku–rusuy	人接－助動：2，他動－助動：5，名詞－助動：0
rusuy–ya	助動－終助：5，助動－名詞：0
ya–?	終助－区切：13，名詞－区切：0

ここで，最も共起頻度の高い品詞列は，次に示すものであり，最も共起頻度の高い2－gramの品詞列と妥当と考えられる品詞列が一致した．

usey　e　= ku　rusuy　ya　?

名詞　人接　　他動　助動　　終助　区切

32　　62　　5　　5　　13　　（アイヌ語品詞列2－gramの共起頻度）

最も共起頻度の高い結果を選択する方略を頻度優先方略と呼ぶことにすれば，ここでの議論は，頻度優先方略によるアイヌ語品詞列の多義性の解消ということができる．

次に，アイヌ語単語「ya」の品詞が終助詞で日本語訳が「か，の，ね」と3通りになる多義性の解消を考えて，階層3で2－gram解析を行うと次のような結果が得られた．

か（終助）－？：10

の（終助）－？：2

ね（終助）－？：1

ここでも頻度優先方略をとると，「ya」の多義性が解消され，単語直接翻訳は次のようになり，妥当な結果が得られる．

usey　e　= ku　rusuy　ya　?

名詞　　人接　　他動　助動　　終助　区切

お湯　　あなた　飲む　たい　　か　　?

名詞　　接代　　他動　助動　　終助　区切

ただし，「の」について，階層3の2－gram解析の結果は，「の」には“ruwe”，“hawe”の形式名詞としての「の」も存在しているため，実際の頻度は「の－？：10」であり，解析に注意が必要である．階層3日本語単語列の2－gramとして，第1層のアイヌ語に依存せず「の－？」の頻度が多いのであれば，頻度優先方略によっては，品詞にも関係なく，「の－？」を

選択することも考えられる．そして，この例では，「の－？」でも自然な日本語と思われる．

（2）複合語の推測

層指定検索ツール[8]と層指定解析ツールを用いて，アイヌ語単語列“somo ki”あるいは“ka somo ki”を複合語として推測する解析過程について述べる．解析のための対訳データは，ALLEU 6 FinalB.txtで，これは文献[1,2]から研究用に作成した階層的な対訳データである．

複合語の推測ということなので，まず，第1層（アイヌ語単語列）で2－gramの解析を行った．その結果の頻度上位10位までのデータを下表に示す．

解析層　1	gram数　2	パターン数　2720	総数　4371
ne–.　：54；ruwe–ne：36；hi–ta　：34；ne–yakka：28；or–ta　：26 pe–ne　：26；somo–ki：24；ku–kor　：23；ka–somo　：20；wa–ku：19			

ここで，頻度24の“somo–ki”と頻度20の“ka–somo”に着目する．これら二つのパターンから“somo”を含む複合語の存在の可能性を推測する．前処理として，層指定検索ツール[8]を用いて“somo”を検索し，“somo”とその同一位置の単位にマーキング処理を行った．これにより，アイヌ語の“somo”を含む対訳データのみを解析することができ，また，マーキング処理により，解析結果を参照するとき“somo”に‘◆’を付加することによって，“somo”と同一位置で発生する語を発見することが可能となる．

次に，このデータを入力ファイルとして層指定解析ツールを利用し，第1層（アイヌ語単語列）と第3層（単語直接翻翻訳）に対して，1－gram，2－gram，3－gramでの解析処理を行った．この結果のパターン数と総数，そして，解析データ内の上位3つのデータをまとめたものが次頁の表である．この解析データにおいて，1－gramのデータから“◆somo◆”（“◆ない◆”）の数は28と確認できる．そのうち，第1層の2－gramでの解析データにおいて，“◆somo◆－ki”（“◆ない◆－する”）が頻度24，“ka－◆somo◆”（“も－◆ない◆”）が頻度20，第1層の3－gramのデータから，“ka－◆somo◆－ki”（“も－◆ない◆－する”）が頻度19と“somo”を含む関係の中で多い．その中で，“somo ki”（“ない　する”）の頻度が24，“ka somo ki”（“も　ない　する”）の頻度が19と多く，複合語としての可能性が高いことが推測される．この推測を元に，層指定検索ツールを利用して，“ka somo ki”を含む対訳データを検索した結果の例を次に示す．

exp1700921：k=eyaysitoma kusu eci=nukare rusuy　◆ka somo ki◆　.

exp1700902：人接＝自動　接助　人接＝複他　助動　副助　副詞　他動　.

exp1700903：私＝はずかしい　から　私があなたに＝見せる　たい　も　ない　する　。

exp1700904：接代＝形容　接助　接代＝他動　助動　副助　助動　他動　。

exp1700905：恥ずかしいから見せたくない。

exp1700906：恥ずかしいから見せたくないなあ。

upa0100501：ne sayo huraruy wa k=eramasu ◆ka somo ki◆ .

upa0100502：連体 名詞 自動 接助 人接＝他動 副助 副詞 他動 .

upa0100503：その おかゆ 匂いが強い て 私＝好む も ない する 。

upa0100504：連体 名詞 形容 接助 接代＝他動 副助 助動 他動 。

upa0100505：そのおかゆは匂いが強くて私は好まない。

upa0100506：そのおかゆは匂いが強くて，私は好きではありませんでした。

この検索結果から，“somo ki”あるいは“ka somo ki”は自然な日本語訳において「～ない」と訳出されるのが妥当な複合語であることが推測できる．

	解析層	gram数	パターン数	総数	解析結果データ：頻度
①	1	1	170	468	.：41 ◆somo◆：28 ka：26
②	1	2	342	443	◆somo◆－ki：24 ka－◆somo◆：20 ki－.：11
③	1	3	379	418	ka－◆somo◆－ki：19 ◆somo◆－ki－.：11 somo–ki–ya：2
④	3	1	180	468	。：41 ◆ない◆：28 する：25
⑤	3	2	355	443	◆ない◆－する：24 も－◆ない◆：20 する－。：11
⑥	3	3	335	418	も－◆ない◆－する：19 ◆ない◆－する－。：11 ない－する－か：2

実際に，文献[5]において，次のような複合語としての取り扱いがなされている．

『 somo ki ：...しない／しなかった

ka somo ki：（動詞句の後で）～しもしない，～したりなんかしない 』

また，文献[7]では，助動詞についての説明の中で，次のような記述が行われている．

『 somo ki 「～しない」

somo kiを助動詞として立てるのはまだ一般的な説ではない．しかし，eaykap「できない」と現れる位置が同じであり，人称変化もしないという点で，somo ki全体を一つの助動詞とみなす根拠があると思われるのでここにあげてみた． 』

4．おわりに

アイヌ語・日本語機械翻訳システムの構築やアイヌ語と日本語の対照言語学的な考察のための基礎的なデータとして，アイヌ語・日本語対訳データは有用である．本報告では，まず，アイヌ語・日本語対訳データの階層的な構成について検討した．次に，階層的な対訳データの階層を指定してN–gram解析を行う層指定解析ツールの構成について述べた．本ツールについては，基本的な解析ツールとしての有効性を確認することができた．一つの階層のN–gram解析だけでなく，二つの階層のN–gram解析の結果を段階的に組み合わせて利用することにより，有用な結果が得られることも示された．アイヌ語・日本語対訳データの作成・蓄積を進めて，層指定解析ツールを利用した統計的な手法の有効性と機械翻訳システム構築における有効な利用法について引き続き検討を進めて行きたい．層指定解析ツールのシステム設計において，オブジェクト指向方法論に基づくUMLモデリングを応用し，設計の効率化とモジュール化を進め，UMLモデリングの有効性を確認することができたこともここに記しておきたい．

謝辞

本研究の一部は，文部科学省「私立大学戦略的研究基盤形成支援事業」による援助を受けて行われました．ここに記して謝意を表します．また，アイヌ語の文法についてご教示をいただいている電子情報工学科切替英雄先生に感謝いたします．

参考文献

1）中川裕，中本ムツ子：エクスプレス　アイヌ語，白水社，1997.
2）中本ムツ子，片山龍峯：アイヌの知恵　ウパシクマⅠ／Ⅱ，片山言語文化研究所，1999／2001.
3）切替英雄：アイヌ神謡集辞典，大学書林，2003.
4）田村すず子：アイヌ語，in「言語学大辞典セレクション：日本列島の言語」，三省堂，1997.
5）田村すず子：アイヌ語沙流方言辞典，草風館，1996.
6）中川裕：アイヌ語千歳方言辞典，草風館，1995.
7）佐藤知己：アイヌ語文法の基礎，大学書林，2008.
8）安曇恭徳・桃内佳雄：層指定検索ツールの開発，工学研究（北海学園大学大学院工学研究科紀要），第8号，pp.53－62，2008.
9）桃内佳雄・大友雄介・越前谷博：アイヌ語・日本語機械翻訳のための基礎的研究，工学研究（北海学園大学大学院工学研究科紀要），第2号，pp.143－151，2002.