

タイトル	アイヌ語地名構成要素解析のための辞書と解析ツールの構成
著者	桃内, 佳雄; 小林, 亮介; Momouchi, Yoshio; Kobayashi, Ryosuke
引用	工学研究 : 北海学園大学大学院工学研究科紀要(10): 39-49
発行日	2010-09-30

アイヌ語地名構成要素解析のための辞書と解析ツールの構成

桃内佳雄*・小林亮介**

Dictionaries and Analysis Tools for the Componential Analysis of Ainu Place Name

Yoshio Momouchi* and Ryosuke Kobayashi**

1. はじめに

アイヌ語地名の構成に関する基礎的な考察と地名構成要素解析の自動化のために、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の構成と電子化は、重要な課題であると考えられる。本報告では、羽田野によってまとめられている異なり語数 231 語の「頻出アイヌ語地名」および北海道環境生活部によって作成されている「アイヌ語地名リスト」に含まれるアイヌ語地名に関するデータに基づく、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の構成について考察する。また、アイヌ語地名構成辞書から構成したアイヌ語日本語地名対訳ファイルを利用してアイヌ語地名に関する N-gram 解析を行うツールを作成し、いくつかの情報について、N-gram 解析実験を行った結果について報告する。

「頻出アイヌ語地名」は、羽田野による「アイヌ語地名の史料における出現頻度」¹⁾でまとめられている。松浦武四郎著(1859年)「東西蝦夷山川地理取調図」、陸地測量部著(1886~1896年測量)「5万分の1地形図」、永田方正著(1891年)「北海道蝦夷語地名解」に現れるすべてのアイヌ語地名の出現頻度の多い順に178位までの異なり語数213語の頻出アイヌ語地名のリストである。「アイヌ語地名リスト」は、平成11年度に北海道が「アイヌ語地名普及会議」を設置し検討を進めた結果としてまとめられた北海道の地名を見出し語とするアイヌ語地名のリストである²⁾。平成19年1月末ま

での改定版が北海道環境生活部のホームページの中で公開されている。

2. アイヌ語地名構成辞書とアイヌ語地名構成要素辞書

2.1 「頻出アイヌ語地名」と「アイヌ語地名リスト」の構成

「頻出アイヌ語地名」のリストは、次のような構成の表としてまとめられている。下表の項目の中で、取調図、仮製図、地名解は、参照している三つの資料に対応している。具体的な例として上位2例を示す。

【順位：地名：ローマ字：取調図：仮製図：地名解：述べ数】

1：ポロナイ：Poronai：45：61：30：136
2：オンネナイ：Onnenai：38：56：27：121

アイヌ語地名は、カタカナ表記とローマ字表記が与えられている。ローマ字表記は、1語として表記されており、その内部の構成要素の区切りは示されていない。従って、このリストを参照して、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書を構成するためには、1語としての地名を構成要素に区切り、その構成要素に対応する日本語での意味を割り当てる作業が必要になる。

「アイヌ語地名リスト」は次のような構成の表と

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

** 北海学園大学工学部電子情報工学科
Dep. of Electronics and Information Eng., Faculty of Eng., Hokkai-Gakuen University

してまとめられている。

【現在の地名；区分；アイヌ語地名〔カナ表記；ローマ字表記〕；アイヌ語の意味；解釈及び由来；出典；備考〔確定レベル；コメント〕】

〔解釈及び由来〕欄があって、〔出典〕に基づく解説が記述されている。また、〔備考〕欄でアイヌ語としての妥当性についての評価の判定を示している。「アイヌ語地名普及会議」での判定の結果がそこにまとめられている。その評価の基準は、『ただし、多くの地名解釈は基本的には推論です。たとえ確定レベルが「A」であっても、断定するものでないことを御理解ください。』という注意を述べて、次のようにまとめている。

『A：次の①～④の全ての条件を満たすもの。

B：少なくとも①及び②の両方を満たすもの。

C：A及びB以外のもの。

- ①音、意味、文法などアイヌ語として適切であると思われる。
- ②地理的条件や事実関係等解釈に妥当性があると思われる。
- ③根拠となった場所や地形、事実等がほぼ確認できる。
- ④従来の諸説に照らし適切であり、また一定程度の定説となっていると思われる。』

地名解析の確からしさを評価する基本的な評価基準がここにまとめられている。このリストの最初の地名を記述例として下に示す。

『・現在の地名（所在地）：アイウシ 愛牛（浦幌町）

・区分：地区

・アイヌ語地名：カナ表記：アイウシニウシイ
*アユシニウシ

・ローマ字表記：ayusi-us-i

・アイヌ語の意味：センの木・群生する・所

・解釈及び由来：{浦幌町郷土博物館報告も同説をとっている。ただし「ayusi」は「ay-us-i（とげ・ついでいる・木）」でタラノキなども考えられる。}

・出典：山田

・備考：確定レベル：B

コメント：①○ ②○ ③ー ④ 』

この資料に含まれているアイヌ語地名に関するデータの中で、本報告における辞書の構成において基本的な情報となるのは、「アイヌ語地名ローマ字表記」と「アイヌ語の意味」である。ローマ字表記に対して、地名の内部の構成要素の区切りと、その区切りに対応する日本語での意味が示されている。この二つのデータを、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の構成における基礎データとして用いることができる。

2.2 アイヌ語地名構成辞書

アイヌ語地名を構成要素へ分解し、その構成要素の意味を日本語への直接翻訳として解釈し、その意味の総和として地名の由来を探るための基本的な辞書として、アイヌ語地名構成辞書の構成を考える。その基本的な構成は次のようである。これは、「アイヌ語地名リスト」の構成を参照し、範疇構成と日本語訳を加える形で構成している。範疇構成は、基本的には品詞構成で、地名の形態論的な構造について考察するとき重要な情報であり、これを加えることにした。例も示す。

【アイヌ語地名：範疇構成：日本語直訳：範疇構成：日本語訳】

【yam-wakka-nai：自動詞＋名詞＋名詞：冷たい|水|川：形容詞＋名詞＋名詞：冷たい水の川】

「アイヌ語地名」は「アイヌ語地名リスト」の「アイヌ語地名ローマ字表記」に対応し、「日本語直訳」は「アイヌ語の意味」に対応している。構成要素の区切りは、アイヌ語地名については“-”，範疇構成については“+”，日本語直訳については“|”として区別する。日本語直訳は、構成要素の直接翻訳をあて、日本語訳は助詞（が、を、に、…）などを補った自然な日本語の訳をあてる。これらを区別したのは、アイヌ語地名の意味をその日本語訳として考えるとき、日本語直訳から日本語訳への変換処理についても考察する必要があると考えているからである。この例では、“冷たい|水|川”から“冷たい水の川”への変換である。アイヌ語地名構成辞書は、地名の構成と構成（要素）に直結した意味（日本語直訳）を含むものとして考えているので、解釈・由来、出典、評価レベルなどに関する情報は含めないものとしている。それらの情報は、「アイヌ語地名リスト」のような構

成で、「アイヌ語地名辞書」の中に記載されており、必要に応じて参照可能となっているものと考えられる。アイヌ語地名構成辞書は、アイヌ語地名辞書の基本的な構成部分という位置づけである。

アイヌ語地名構成辞書をこのような形でまとめることにはいくつかの問題点がある。

- (1)アイヌ語地名の構成：必ずしも一意に構成が決まるとは限らない。
- (2)アイヌ語地名構成要素の多義性：異なる日本語直訳が対応する場合がある。
- (3)範疇（品詞）の種類：アイヌ語での品詞の曖昧さ（連体詞か接頭辞か、場所の意味を含む名詞か位置名詞か、など）、アイヌ語品詞と日本語品詞の対応関係の一意性などの問題についても検討しなければならない。
- (4)日本語直訳：1語の単語が対応するとは限らない。「アイヌ語地名リスト」では、格助詞が付与されている例や、句、節、文、文章などが対応している例がある。文、文章になると日本語直訳というよりは、意味の記述である。この場合の範疇は品詞の枠を越えている。

2.3 アイヌ語地名構成要素辞書

アイヌ語地名構成要素辞書は、アイヌ語地名を構成する要素を収める辞書である。アイヌ語地名を構成する要素にはどのようなものがあり、また、それがどのように組み合わせられて地名が構成されるかなどの解析のための基本的な辞書となる。その構成を次のように考える。

【地名構成要素：アイヌ語範疇：日本語：日本語
範疇：補足情報（方言・地方）】

補足情報を除く基本的な部分は、前節で構成したアイヌ語地名構成辞書から自動的に構成される。「日本語」は、基本的には、アイヌ語地名構成要素の日本語直訳であるが、対応する日本語直訳がなく、その範疇が「記述」となる場合がある。「補足情報」は、例えば、「方言」、「地方」などに関する情報をここに載せる。これは、アイヌ語地名辞書の構成を前提として、そこに記載されている情報を利用することを考えている。知里の「地名アイヌ語小辞典」³⁾には方言情報も記載されている。例えば、この情報を利用して、異なる方言の構成要素の組み合わせを持ったアイヌ語地名の構成を

チェックすることができる。具体例を示す。

- ①【yam：自動詞：冷たい：形容詞：H北；K】
- ②【wakka：名詞：水：名詞： 】
- ③【nai：名詞：川：名詞： 】
- ④【ununkoy：名詞：川の両岸が狭い断崖になっていて、川伝いに登って行った人が、そこから先へは通りぬけることができず引き返さねばならぬような地形：記述：】

①の例では、アイヌ語と日本語の品詞が異なっている。補足情報の「H北」は北海道北部方言地帯、「K」は樺太である(知里³⁾)。④の例は、日本語が記述の例(知里³⁾)である。アイヌ語の品詞として名詞が妥当かどうか検討の余地がある。

アイヌ語地名構成要素解析のための基本的な辞書として、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の構成について検討した。従来のアイヌ語地名辞典、アイヌ語地名解、アイヌ語地名リストに含まれているアイヌ語地名に関する情報を整理し、また、アイヌ語地名構成要素解析の計算機による自動化の可能性を考慮して、その基本的な構成について検討した。

2.4 アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の作成

アイヌ語地名構成辞書の最初の要素として、その構成が解析ずみのアイヌ語地名を置くこととしている。その構成の解析をどのように行うかが問題である。最初の出発点は、人間が解析するところから始まる。そして、その解析の結果を用いて、アイヌ語地名構成要素辞書が作られていけば、その利用が構成の解析を助けることになる。このような段階的な構成と解析のサイクルによって、少しずつ、アイヌ語地名構成辞書、アイヌ語地名構成要素辞書が構築されてゆくというプロセスを考えることができる。そのようなプロセスの基盤として、切替による「頻出アイヌ語地名の形態論的構造」についての詳細な考察⁴⁾と「アイヌ語地名リスト」のアイヌ語地名ローマ字表記を置くこととした。それらを基礎データとして、アイヌ語地名構成辞書とアイヌ語地名構成要素辞書の作成を進めている。まず、この二つのリストから、アイヌ語地名構成辞書を作成する手順をまとめる。

(1)「頻出アイヌ語地名の形態論的構造」⁴⁾に基づきアイヌ語地名構成辞書を作成する手順

- ①アイヌ語地名ローマ字を構成要素に分割する。
- ②構成要素の範疇（品詞）を割り当てる。
- ③日本語直訳を決める。
- ④日本語直訳の範疇（品詞）を割り当てる。
- ⑤日本語訳を割り当てる。

「頻出アイヌ語地名」については、切替による形態論的構造に関する考察を参照しつつ、構成辞書の作成を人手で進めている。例えば、切替⁴⁾では、「poromoi」という地名について次のような解析を行っている。文献¹⁾では、「Poromoi」と先頭の文字が大文字で表記されているが、ここでは、小文字で表記することにする。

『poromoi 40 23 moi 大きな入り江；構成：「自動詞＋主語相当語」』

「40 23」は「頻出アイヌ語地名」における順位と頻度である。「moi：入り江」という解析により、poromoiが、「poro-moi」という構成で、日本語直訳「大きい | 入り江」、日本語訳「大きい入り江」という解析が可能となる。ここでは、辞典³⁾を参照して、poroの日本語直訳は「大きい」とした。品詞情報については、アイヌ語については、「自動詞＋名詞」、日本語については「形容詞＋名詞」という解析が可能となる。

【poro-moi：自動詞＋名詞：大きい | 入り江：形容詞＋名詞：大きい入り江】

このように、切替⁴⁾による解析には、「頻出アイヌ語地名」に含まれている地名の構成辞書を作成するための基本的で、重要な情報が含まれており、これを参照しながら、また辞典^{3,14,15)}も参照しながら、「頻出アイヌ語地名リスト」に対するアイヌ語地名構成辞書の作成を進めている。

(2)「アイヌ語地名リスト」²⁾に基づきアイヌ語地名構成辞書を作成する手順

- ①アイヌ語地名ローマ字表記は構成要素に分割されているので、そのままの構成とする。
- ②構成要素の範疇（品詞）を割り当てる。
- ③日本語直訳は構成要素に分割されているので、原則として、そのままの構成とする。一つの単語が対応していない場合にも、まず第1近似としてそのままの構成とする。
- ④日本語直訳の範疇（品詞）を割り当てる。
- ⑤日本語訳を割り当てる。

先にも述べたように、「アイヌ語地名リスト」で

は、「アイヌ語地名ローマ字表記」が「アイヌ語地名」に直接対応し、「日本語直訳」は「アイヌ語の意味」に直接対応している。範疇（品詞）情報については、「頻出アイヌ語地名」に対する(1)の作成過程で得られた情報や、辞典^{3,14,15)}を参照しながら、割り当てていく。

アイヌ語地名構成辞書が作成できれば、アイヌ語地名構成要素辞書は、半自動的に構成することができる。アイヌ語地名構成辞書の一つの項目が次のように構成されているとすると、

【yam-wakka-nai：自動詞＋名詞＋名詞：冷たい | 水 | 川：形容詞＋名詞＋名詞：冷たい水の川】

区切り記号（“-”，“|”，“+”）を手がかりとして、構成要素辞書の始めの4つの項目を自動的に切り出すことができる。例えば、上例からは次のような項目を切り出すことができる。この構成での構成要素辞書を基本構成要素辞書と呼ぶ。

【yam：自動詞：冷たい：形容詞： 】

【wakka：名詞：水：名詞： 】

【nai：名詞：川：名詞： 】

これに第5番目の項目である補足情報を適宜加えることで、構成要素の項目をすべて作ることができる。アイヌ語地名構成辞書から基本構成要素辞書を作成するツールも開発している。構成されているアイヌ語地名構成辞書とアイヌ語地名基本構成要素辞書の一部を以下に示す。

・アイヌ語地名構成辞書（「アイヌ語地名リスト」）

ayusni-us-i：名詞＋他動詞＋接尾辞：センの木 | 群生する | 所：名詞＋他動詞＋名詞：センの木が群生する所

aykap：自動詞：できない：複合語：できない

aynu-oma-nay：名詞＋他動詞＋名詞：アイヌ | いる | 沢：名詞＋自動詞＋名詞：アイヌがいる沢

aynu-o-nay：名詞＋他動詞＋名詞：アイヌ | いる | 沢：名詞＋自動詞＋名詞：アイヌがいる沢

ay-pet：名詞＋名詞：矢 | 川：名詞＋名詞：矢の川

hure-pet：自動詞＋名詞：赤い | 川：形容詞＋名詞：赤い川

hure-suma：自動詞＋名詞：赤い | 石：形容詞＋名詞：赤い石

・アイヌ語地名基本構成要素辞書

a：他動詞：座る：自動詞：
a：名詞：我ら：名詞：
ahun：他動詞：入る：他動詞：
ak：自動詞：浅い：形容詞：
akam：名詞：車輪：名詞：
akkesi：名詞：牡蠣：名詞：
ama：他動詞：置く：他動詞：
amam：名詞：穀物：名詞：
an：名詞：山むこう：名詞：
an：名詞：片側：名詞：
an：名詞：鷺捕りの雪穴あるいは小屋：名詞：

2.5 アイヌ語地名解析のための基本的な辞書

アイヌ語地名解析のための基本的な辞書とその役割について、ここで整理する。

(1) アイヌ語地名辞書

アイヌ語の地名を見出し語とする辞書で、「アイヌ語地名リスト」の内容に対応する様々な情報を蓄積する。見出しは、原則として、構成要素ごとに区切りをいれたローマ字表記とする。対応する日本語直訳も区切りをいれた表記とする。「アイヌ語地名リスト」を基盤として作成することができる。アイヌ語地名についての基本辞書である。切替⁴⁾でも解析しているように構造のわからない地名もあるということを考慮して、見出しは区切りなしの表記も可とし、その旨の印をつける。

(2) アイヌ語地名構成辞書

アイヌ語地名辞書から、その骨格部分を抽出し、範疇(品詞)構成を追加した辞書である。アイヌ語地名の構成パターン・形態論的構造の考察と解析のための基礎的な辞書となる。

(3) アイヌ語地名構成要素辞書

アイヌ語地名構成辞書から、構成要素を切り出し、構成要素を見出しとする辞書である。補足情報として、方言・地方情報を付加する。アイヌ語地名解析のための基礎的な辞書となる。方言・地方情報を参照することにより、地名解析の精度を上げることができる。

(2)と(3)について、本章で、具体的な構成と基本的な作成手順について考察を行った。

3. アイヌ語地名の形態論的構造

切替⁴⁾は、「頻出アイヌ語地名の形態論的構造」を下記のようにまとめている。地名の基本的構成は、[<修飾部>+<主要部>]と考えることができ、(1)~(4)を基本的な構成とし、さらに、<修飾部>の構造と<主要部>の有無に着目した分類となっている。

『(1) 名詞

(2) 名詞+名詞

(3) 名詞+位置名詞

(4) 連体詞+名詞

(5) 修飾構造をなす地名

(5.1) 自動詞+主語相当語

(5.2) 主語+他動詞+目的語相当語

(5.2.1) 目的語相当語の位置を名詞的接尾辞が占めているもの

(5.3) 目的語+他動詞+主語相当語

(5.3.1) 主語相当語の位置を名詞的接尾辞が占めているもの

(5.4) 主語(所属物)+自動詞+斜格語相当語(所属先)

(5.5) 主語(所属物)の位置を部分接頭辞が占めているもの

(5.5.1) 主語相当辞(所属物)+目的語+他動詞+斜格語相当辞(所属先)

(5.6) 目的語相当辞(所属物)+主語+他動詞+斜格語相当語(所属先)

(5.6.1) 斜格語相当語(所属先)の位置を名詞的接尾辞が占めているもの

(6) 疑似修飾構造をなす地名

(6.1) 自動詞+ush+斜格語相当語

(6.1.1) 斜格語相当語の位置を名詞的接尾辞が占めているもの

(6.2) 目的語+他動詞+ush+斜格語相当語

(6.2.1) 斜格語相当語の位置を名詞的接尾辞が占めているもの

(6.3) 他動詞+ush+斜格語相当辞

(7) 主要部のない地名

(7.1) 自動詞

(7.2) 主語+他動詞

(7.2.1) 不定人称接頭辞主格形が主語の位置を占めているもの

(8) 構造のわからない地名

』

修飾構造と疑似修飾構造は連体節修飾句を分類したものである^{5,6)}。また、斜格語相当語(辞)は「の格」に対応する語(辞)である。疑似修飾構造をなす地名は、“ush”を間に挟む構造のみがまとめられている。この分類は、「頻出アイヌ語地名」に含まれるアイヌ語地名について試みたものであるが、この分類を基礎として、若干の拡張を加えた構成についても考察を進めることができる⁷⁾。また、アイヌ語地名を構成する基本構成要素は、単語と接辞(接頭辞、接尾辞)であることがわかる。上の分類と考察⁷⁾に基づいて、品詞範疇の並びに着目したアイヌ語地名の基本的なパターンを次のようにまとめることができる。「/」は「または」を意味する。

- ①名詞
- ②名詞+名詞
- ③名詞+位置名詞
- ④連体詞+名詞
- ⑤自動詞+名詞(/ 接辞)
- ⑥名詞(/ 接辞)+他動詞+名詞/接辞
- ⑦名詞/接辞+自動詞+名詞(/ 接辞)
- ⑧接辞+名詞+他動詞+名詞/接辞
- ⑨自動詞+ush+名詞/接辞
- ⑩名詞(/ 接辞)+他動詞+ush+名詞/接辞
- ⑪他動詞+ush+接辞
- ⑫自動詞
- ⑬名詞+自動詞
- ⑭名詞/接辞+他動詞

構成における前方位置の接辞は接頭辞、後方位置の接辞は接尾辞である。接頭辞は、o(尻)、接尾辞は、pe(形式名詞¹⁴⁾:もの、者、こと)、i(もの、ところ、とき、こと)、p(もの)などである。これらはすべて名詞的接辞である。「(/接辞)」は、文献⁷⁾での考察に基づいて追加されたパターンである。上の構成は基本的なパターンで、これらの構成、特に①から⑧の構成を名詞句として入れ子とする構造のパターンも可能で、切替⁴⁾には、いくつかの階層的な構成の例が示されている。

- 名詞句+名詞
Yam-wakka-nai :
[自動詞+名詞]+名詞
[冷たい | 水] | 川 : 冷たい水の川
- 名詞句(目的語)+他動詞+名詞

Nupuri-pa-oma-nai :

[名詞+名詞]+他動詞+名詞

[山 | 上手] | ある | 川 : 山の上手にある川

• 自動詞+名詞句

Poropi-nai :

自動詞+[名詞+名詞]

大きい | [石 | 沢] : 大きい石沢

すぐ上の例は、「[自動詞+名詞]+名詞」という構成も可能であるように思われる。アイヌ語地名構成辞書における範疇構成では、このような階層的な構成は記述しない。階層的な構成の解析は、より進んだ処理として考えることとしたい。

4. アイヌ語地名構成要素解析のための N-gram 解析ツール

アイヌ語地名構成要素解析のための基本的な解析ツールとして、N-gram 解析ツールを作成した。アイヌ語地名構成要素解析にとって基本的な情報である、アイヌ語地名、品詞範疇構成、日本語直訳に着目して N-gram 解析を行うツールである。アイヌ語地名構成辞書の構成を進めながら、一方で、この三つの構成要素のみからなるファイルを地名対訳ファイルとして構成し、アイヌ語地名構成要素解析に関する基礎的な実験を行っている。

4.1 地名対訳ファイル

地名対訳ファイルの構成は次のようである。

【アイヌ語地名：範疇構成：日本語直訳】

このような構成の地名対訳ファイルを「頻出アイヌ語地名」と「アイヌ語地名リスト」について、アイヌ語地名構成辞書を作成する過程の中で構成している。それぞれの例を下に示す。

[頻出アイヌ語地名]

kamui-chise : 名詞+名詞 : 神 家 poro-nai : 自動詞+名詞 : 大きい 川 : hup-ush-nai : 名詞+他動詞+名詞 : トドマツ ある 川 :
--

[アイヌ語地名リスト]

ayusni-us-i：名詞＋他動詞＋接尾辞：センの
木 | 群生する | 所：

aynu-oma-nay：名詞＋他動詞＋名詞：アイヌ
 | いる | 沢：

現時点で、それぞれのファイルと2つのファイルをあわせたファイルに含まれている異なりアイヌ語地名の数は次のようである。ただし、「nai」と「nay」、「-s」と「-sh」などの異表記は異なるものとして計数している。

- ・「頻出アイヌ語地名」対訳ファイル：213
- ・「アイヌ語地名リスト」対訳ファイル：1063
- ・「頻出アイヌ語地名＋アイヌ語地名リスト」対訳ファイル：1246

また、以下の節で解析ツールの対象となる対訳ファイルとそのデータ数は次のようである。

- ・「頻出アイヌ語地名」対訳ファイル (H4v1)：213
- ・「アイヌ語地名リスト」対訳ファイル (AL18v1)：1090
- ・「H4v1, AL18v1の合成」対訳ファイル (H4AL18v1)：1283

ただし、これらの対訳ファイルは、上述の異表記の問題なども含めて、その内容について、現在検証中の段階にあり、より良いものに改良を進めていくことが今後の課題である。

4.2 N-gram 解析ツール

N-gram 解析ツールは、地名対訳ファイルをデータとして、アイヌ語地名の構成に関する N-gram 解析に加えて、対応する品詞と日本語の構成に関する N-gram 解析も行うことができる。

- ・基本 N-gram 解析
 - ・アイヌ語 (地名構成要素の N-gram 解析)
 - ・品詞 (品詞の N-gram 解析)
 - ・日本語 (日本語直訳の N-gram 解析)

基本 N-gram 解析は、アイヌ語地名に対しては N 個の構成要素の並び、品詞に対しては N 個の品詞の並び、日本語に対しては N 個の日本語直訳の並びの切り出しとその頻度の計数を行う。また、アイヌ語、日本語、品詞を二つ組み合わせた N-gram 解析も可能である。

- ・組み合わせ N-gram 解析
 - ・アイヌ語＋日本語
 - ・アイヌ語＋品詞
 - ・品詞＋日本語

N-gram 解析ツールの基本的な実行画面を図 1 に示す。「頻出アイヌ語地名」対訳ファイルについて、アイヌ語地名構成要素の基本 1-gram 解析を行った結果を示している。「地名対訳ファイル」欄に対訳ファイル名 (H4v1.txt)、解析結果を格納するファイル名 (H4v1Out1.txt) を「N-gram のファイル」欄に入力する。「gram 数」欄には 1 を入力している。解析はアイヌ語地名について行い、その構成要素の区切り記号は「-」であることを入力している。結果は構成要素の辞書引き順に、頻度を添えて出力される。右下の表示欄には、1-gram の全パターン数が 537 個でそのうち異なるパターン数が 160 個であることが示されている。その結果を頻度順ソートツールで処理した結果が図 2 である。「nai」の頻度が多いことが分かる。

次の図 3 は組み合わせ N-gram 解析の実行例である。アイヌ語と日本語の組み合わせについて、1-gram 解析を行った結果が示されている。これを頻度順ソートツールにより処理した結果が図 4 である。「nai：川」の組み合わせ頻度が最も多く 66 となっている。10 番目に「nai：沢」が頻度 8 として出力されている。「nai」を「川」または「沢」のどちらに直訳するかは、文脈 (状況) による。切替⁴⁾では、『pet と nai はあえて訳さずペツ、ナイとした。両者とも「川」と訳されることが多いが、



図 1 N-gram 解析ツールの実行画面

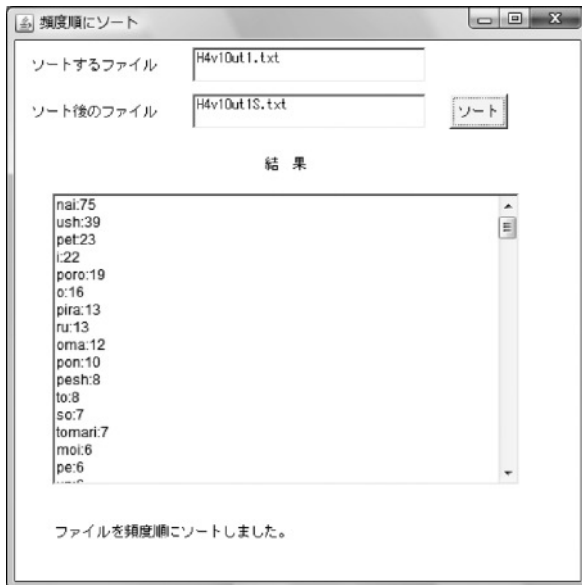


図 2 1-gram 解析結果の頻度順ソート

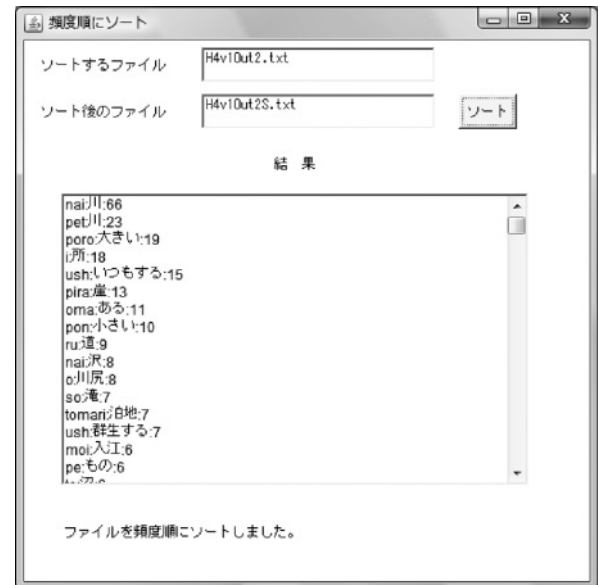


図 4 組み合わせ解析結果の頻度順ソート

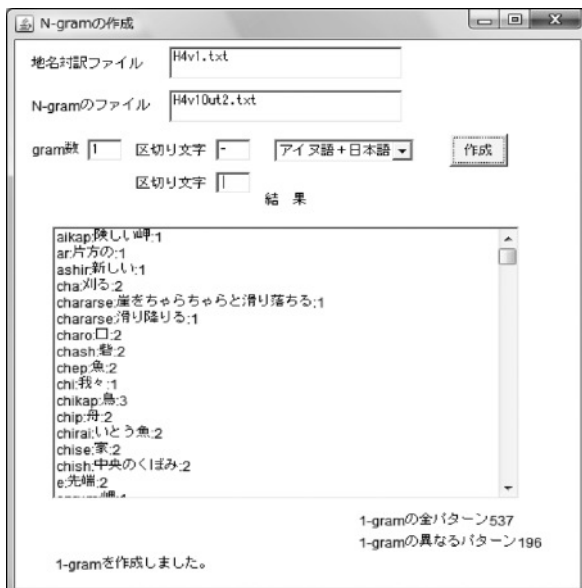


図 3 組み合わせ 1-gram 解析の実行例

ナイはまた「沢」と訳されることもある。」としている。知里³⁾あるいは山田⁸⁾からの引用として示されている訳では、「川」あるいは「沢」があげられている。

「頻出アイヌ語地名」の構成要素辞書は、切替⁴⁾を参考にして作成し、ナイとしている場合の日本語直訳には「川」をあげているので、上の結果にはこのことが影響している。「頻出アイヌ語地名」に現れる「nai」と切替⁴⁾における対応を次の表 1 にまとめる。[出典]は、切替で(C)知里³⁾、(Y)山田⁸⁾として引用されているものからである。[出典]では、「nai:川」が多い。「アイヌ語地名リス

表 1 「頻出アイヌ語地名」における「nai」

nai の出現総数	75
切替 ナイ	39
[出典] 川	26
[出典] 沢	9
[出典] 川, 沢	1

ト」での解析でも、「nai:川」が 123、「nai:沢」が 27 と「nai:川」が多い。

「アイヌ語地名リスト」について、組み合わせ(アイヌ語+日本語)1-gram 解析を行い、その結果をソートしたものが図 5 である。「pet:川」が 232 と多いが、「頻出アイヌ語地名」とは「pet」と「nai(nay)」の多さが逆になっている。

以上のように、本研究で開発した N-gram 解析ツールは、地名対訳ファイルをデータとして、アイヌ語地名の構成要素や構成要素間の関係についての数量的な情報を解析することができる。次章では一歩進んだ解析例について検討する。

なお、N-gram 解析のために開発したプログラムは、本章で利用した N-gram 解析ツール、頻度順ソートツール、次章で利用する N 限定頻度解析ツールの三つであり、プログラミング言語 Java により実現している。

5. N-gram 解析ツールを用いたアイヌ語地名構成要素と構成パターンの解析

N-gram 解析ツールを用いて、「頻出アイヌ語地

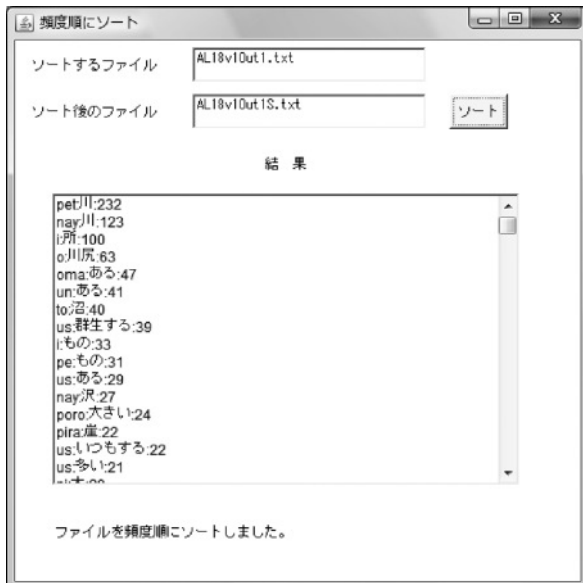


図5 「アイヌ語地名リスト」組み合わせ(アイヌ語+日本語) 1-gram 解析結果の頻度順ソート

名」と「アイヌ語地名リスト」対訳ファイルの合成ファイルについて、アイヌ語地名の構成要素の1-gram 解析を行った結果を図6に示す。全構成要素数が3260個でそのうち異なり数が630個であることが示されている。630個の異なる構成要素が解析されたということである。ただし、この合成ファイルは、二つのファイルに含まれていた、まったく同じデータ要素は一つに縮約しているが、一字(記号)でも異なるデータ要素はそのまま残している。前章でも述べたように、そのようなデータ要素の詳細な分析は今後の課題とした

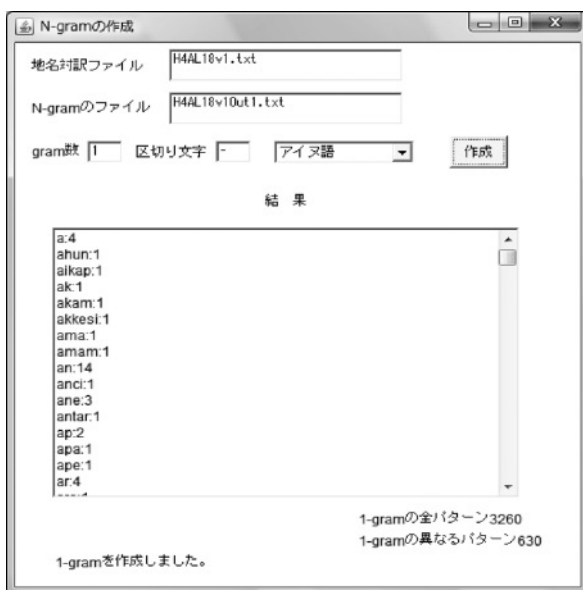


図6 合成ファイルの1-gram 解析の結果

い。

図7は合成ファイルの1-gram 解析結果の頻度順ソートである。この結果における上位5位以内のアイヌ語地名構成要素が頻度100以上であることが分かる。アイヌ語地名構成要素として、「pet」(日本語直訳「川」)が特に多いことが数値データとして示された。「nay」、「nai」も日本語直訳として、「沢」あるいは「川」が対応しており、合わせて頻度474個となる。

表2は上位30位までの構成要素のリストであ

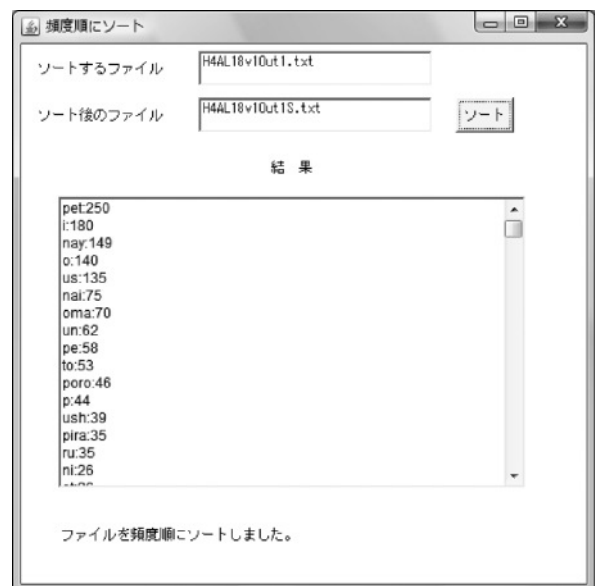


図7 1-gram 解析結果の頻度順ソート

表2 構成要素の頻度30位までのリスト
(「頻出アイヌ語地名」+「アイヌ語地名リスト」)

構成要素	頻度	構成要素	頻度
pet	250	ni	26
i	180	ot	26
nay	149	ota	26
o	140	sir	26
us	135	e	25
nai	75	or	25
oma	70	kotan	22
un	62	panke	22
pe	58	sar	22
to	53	kot	20
poro	46	kus	20
p	44	so	20
ush	39	mo	19
pira	35	penke	19
ru	35	pon	19

る。また、前にも述べたように、「nay」と「nai」、
「ush」と「us」は同義異表記である。対訳ファイルは、現時点で検証中であり、表 2 の頻度は、若干の変化があるかもしれない。およその傾向として捉えたい。

次に、アイヌ語地名の構成について調べるために、品詞について、N-gram 解析を行ってみよう。対象とする対訳ファイルは「頻出アイヌ語地名＋アイヌ語地名リスト」合成対訳ファイルである。N-gram 解析における全パターン数と異なりパターン数を表 3 にまとめる。

1-gram 解析における異なりパターン数 13 は、構成要素の品詞の数に対応する。また、構成要素の数が 6 以上の地名はなかった。2-gram 解析の頻度順ソート結果を図 8 に示す。「名詞＋名詞」、「名詞＋他動詞」、「他動詞＋名詞」、「自動詞＋名詞」、「他動詞＋接尾辞」というパターンが 100 以上で、多いことが分かる。ただし、この 2-gram は、長さがちょうど 2 の構成ということではない。長さが最大 5 の地名の中での 2-gram のパターンを解析した結果である。

ここで、構成要素数 N のデータだけを取り出して頻度の計数を行う N 限定頻度解析ツールを作成して、構成要素の数がちょうど 2 の地名について、その範疇（品詞）構成を解析した結果、全パターン数 554、異なり数 18 という結果が得られ、その結果をソートしたものが図 9 である。この結果は、「頻出アイヌ語地名＋アイヌ語地名リスト」合成対訳ファイルについてではあるが、2 個の要素から構成されるアイヌ語地名の品詞構成が、18 パターンあり、「名詞＋名詞」(268 個)、「自動詞＋名詞」(160 個)が多いということがわかる。3 章でまとめたパターンの②と⑤に対応している。

このような品詞 N 限定頻度解析によって、アイヌ語地名の品詞構成パターンの出現頻度の解析を行うことができ、3 章での考察を確認し、拡張することができる。N が 1 から 5 限定の解析結果を表 4 にまとめている。

表 3 品詞の N-gram 解析

gram 数	全パターン数	異なりパターン数
1	3260	13
2	1977	47
3	807	84
4	189	57
5	10	10

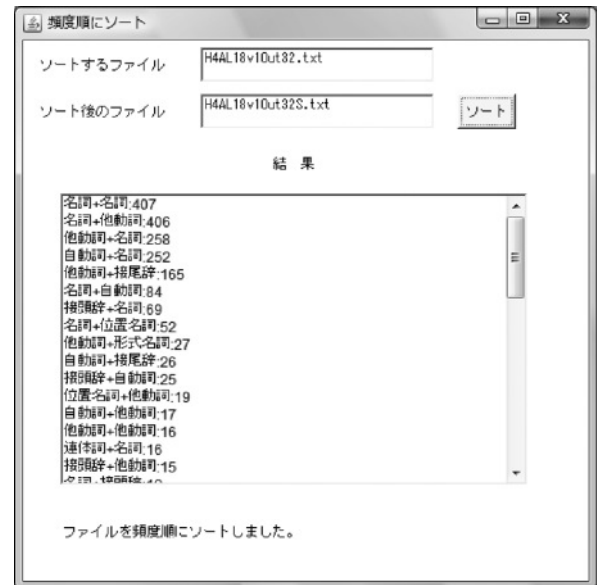


図 8 品詞 2-gram 解析の結果



図 9 品詞 2 限定頻度解析の結果

表 4 品詞構成パターンの頻度解析結果

N	全頻度 異なり数	頻度 2 位まで	頻度
1	112	名	101
	3	自動	7
2	554	名＋名	268
	18	自動＋名	160
3	438	名＋他動＋名	152
	49	名＋他動＋接尾	91
4	169	名＋名＋他動＋名	22
	45	名＋名＋他動＋接尾	19
5	10	10 個全部異なる	
	10		

6. おわりに

「頻出アイヌ語地名」と「アイヌ語地名リスト」に基づいて構成中のアイヌ語地名構成辞書とアイヌ語地名構成要素辞書, そして対訳ファイルの整備と検証を進めること, および, 文献^{12,13)}などを参照して, データを増やしていくことが今後の課題である。また, アイヌ語地名解析システムに N-gram 解析の結果である数量的な(統計的な)情報を組み込んで, 解析の可能性と精度を高める仕組みについても考察を進めているところである。

謝辞

本研究の一部は, 北海学園大学ハイテク・リサーチ・センター研究費による援助を受けて行われました。また, アイヌ語の文献と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) 羽田野正隆: アイヌ語地名の史料における出現頻度, 北方文化研究, No.20, pp.17-32, 1989.
- 2) 北海道環境生活部: アイヌ語地名リスト, (財)アイヌ文化振興・研究推進機構, 2004.
(<http://www.pref.hokkaido.lg.jp/ks/sum/soumuka/ainu/ainurist.htm>, 改訂版, 2007)
- 3) 知里真志保: 「地名アイヌ語小辞典」, 北海道出版企画センター, 2000 (復刻六刷; 初版 1956).
- 4) 切替英雄: 頻出アイヌ語地名の形態論的構造, アイヌ語地名研究, No.3, pp.105-142, 2000.
- 5) 切替英雄: アイヌ語の名詞句の構造と合成名詞, 言語研究, No.86, pp.105-121, 1984.
- 6) 桃内佳雄: アイヌ語と日本語の連体節修飾名詞句の基本的な構成と対訳パターン, 北海学園大学工学部研究報告, 第 32 号, pp.181-202, 2005.
- 7) 桃内佳雄: アイヌ語地名解析の自動化のための基礎的考察, 北海学園大学工学部研究報告, 第 33 号, pp.153-169, 2006.
- 8) 山田秀三: 「アイヌ語地名の研究 (全 4 巻)」, 草風館, 2000 (復刻版; 初版 1982).
- 9) 山田秀三: 「北海道の地名: アイヌ語地名の研究別巻」草風館, 2000 (復刻版; 初版 1984).
- 10) 山田秀三: アイヌ語地名を歩く, 北海道新聞社, 1986.
- 11) 切替英雄: 山田秀三のアイヌ語地名研究, 北海道立アイヌ民族文化研究センター研究紀要, pp.200-218, 2005.
- 12) 永田方正: 「北海道蝦夷語地名解」, 国書刊行会, 復刻

版, 1972.

13) 由良勇: 上川郡内石狩川本支流アイヌ語地名解, 北海道出版企画センター, 2004.

14) 田村すず子: アイヌ語沙流方言辞典, 草風館, 1996.

15) 中川裕: アイヌ語千歳方言辞典, 草風館, 1995.