

音声記号等で表記された言語資料の
マークアップとコンピュータ処理

**Annotation and Computer Processing of Language Resources
in Non-Latin Scripts and Phonetic Transcription**

課題番号 15202008
Grant No. 15202008

平成15年度～平成17年度

科学研究費補助金基盤研究(A)研究成果報告書

2003–2005 Grants-in-Aid for Scientific Research (A) Research Report

研究成果報告書
Research Report

平成 18 年 3 月
March 2006

研究代表者 松村 一 登
(東京大学大学院人文社会系研究科教授)

Project Leader: Kazuto Matsumura
(Graduate School of Humanities and Sociology, University of Tokyo)

アバール語のテキストの電子化

山田久就

小樽商科大学

hisanari@res.otaru-uc.ac.jp

1. はじめに

本稿の目的は筆者がこれまでに取り組んできたアバール語のテキストの電子化について報告することである。

アバール語は北東コーカサス諸語(または、ダゲスタン諸語とも呼ばれている)に属し、主にロシア連邦ダゲスタン共和国およびその南にあり旧ソ連からの独立国であるアゼルバイジャン共和国で話されている。アバール語の標準語は文語を持っていて、キリル文字体系を用いて表記される。

筆者がこれまでに取り組んできたのはアバール語の標準語およびヒダス方言のテキストの電子化である。アバール語の標準語はフンザフ方言を基礎としている。ただし、現在において標準語とフンザフ方言には語彙においても文法においても十分な違いがある。また、アバール語は山岳地域で話されていることなどから方言差が大きく、標準語も使用者の出身地などによってかなりのゆれがある。一方、ヒダス方言と標準語にはかなり大きな違いがある。ただし、ヒダス方言より標準語との違いが大きい方言も存在する¹。

標準語のテキストは本からのテキストであり、スキャナーとOCRを使用して電子化を行った。一方、ヒダス方言には文語がなく、ヒダス方言の話者に依頼して作文してもらったテキストである。

テキストの電子化の主要な目的は、筆者自身が行っているアバール語の研究を効率的に行うためである。しかし、電子化を行ったテキストを共同利用できるように公開することを目指している。そのため、ヒダス方言の母語話者にテキストの作成を依頼するにあたっては一般公開の許可を得ている。

2. 文字と句読点

2.1. 標準語

アバール語の標準語ではキリル文字体系の一種を使用しているが、使用されている具体的な文字は次の46文字である。

1	2	3	4	5	6	7	8	9	10
а	б	в	г	гъ	гь	гІ	д	е	ё
[a]	[b]	[w]	[g]	[ɣ]	[h]	[ʕ]	[d]	[je], [e]	[jo]
11	12	13	14	15	16	17	18	19	20
ж	з	и	й	к	къ	кь	кІ	л	лъ
[z]	[z]	[n]	[j]	[k]	[kχ']	[tʃ']	[k']	[l]	[ɫ]
21	22	23	24	25	26	27	28	29	30
м	н	о	п	р	с	т	тІ	у	ф
[m]	[n]	[o]	[p]	[r]	[s]	[t]	[t]	[u]	[f]
31	32	33	34	35	36	37	38	39	40
х	хь	хь	хІ	ц	цІ	ч	чІ	ш	щ
[χ]	[kχ]	[ç]	[h]	[ts]	[ts']	[tʃ]	[tʃ']	[ʃ]	[ʃʃ]
41	42	43	44	45	46				
ъ	ы	ь	э	ю	я				
[ʔ]			[e]	[ju]	[ja]				

表1 文字

42番目のыおよび43番目のьはロシア語など他の言語からの借用語でしか使われない。アバール語にはロシア語からの借用語がとてまたくさんあるが、ロシア語からの借用語はアバール語の発音とは関係なしにロシア語の綴りのままで表記することになっている。

5番目のгъはгとъという二つの部分(以下では便宜的に字素と呼ぶ)からできている。このように、アバール語で使われる文字には二つの字素からできているものがあり、アバール語で使われている字素は全部で34個である。また、10番目の文字ёはйоで表記することが一般的であり、この文字を使用している文献は少ない。

アバール語にはハイフン(-)を含んだ語が多くみられる。そのような語には二つの語を連続的に並べた複合語が多い。たとえば、гъава-бакъ「天気」はгъава「空気」とбакъ「太陽」からの複合語である。

アバール語で使われている句読点は次の通りである。

1	2	3	4	5	6	7	8	9	10
.	,	?	!	:	;	()	«	»
11	12	13							
“	”	—							

表2 句読点

9, 10 番目の « と » および 11, 12 番目の “ と ” はともに引用符として用いられるのが普通で、一つの文献ではどちらか一方が用いられる。9, 10 番目の « と » が用いられることの方が多い。13 番目の横棒(—)はいろいろな場合に用いられるが、会話の引用の開始(あるいは終了)を示すのに用いられることが多い。

こうした文字(あるいは字素)と句読点で書かれているアバール語の標準語のテキストを電子化するために用いた文字コードはユニコード(Unicode)系の utf-8 である²。アバール語で用いられているのはキリル文字体系の一種であり、キリル文字体系を使った言語を表示する際に使われる文字コードには cp1251、MacCyrillic、koi8-r、cp866 などがあるが、将来的な拡張性を考えてユニコード系の utf-8 を用いている。ちなみに、cp1251 は Windows で、MacCyrillic は Macintosh で、koi8-r は Unix 系 OS でロシア語をはじめとするキリル文字体系を使う言語を表示するために標準的に使われている文字コードである。また、koi8-r はロシア語で e-mail をやりとりする際にも標準的に使われている。cp866 は DOS でロシア語などのキリル文字体系を使う言語を表示するためによく使われた文字コードである。

ユニコードではアバール語を含む北コーカサス諸語の字素 I を表示するためにコード番号 U+04C0 が用意されているが、この U+04C0 を使わずにコード番号 U+0406 を用いている。U+0406 はウクライナ語などの I を表示するための文字コードである。U+0406 を用いている理由は北コーカサス諸語の字素 I を表示するための文字コード U+04C0 のグリフを欠いているフォントが多いことである。また、ウクライナ語などの I はアバール語を電子化する限りにおいて今後も使う可能性がないという理由もある。ラテン文字体系の I もほぼ同じ見目をしているが、ラテン文字体系の I はテキストで本来の目的のために使われているので、両者の混乱が生ずるために利用していない。他の文字コードとの変更可能性について述べると、ウクライナ語の I を表示するためのコード番号は cp1251、MacCyrillic には用意されているが、koi8-r、cp866 には用意されていない。北コーカサス諸語の I を表すためのコード番号は cp1251、MacCyrillic、koi8-r、cp866 のどれにも用意されていない。

ハイフンにはコード番号 U+002D(HYPHEN_MINUS)を用いている。ユニコードではハイフンとマイナスを区別できるようにハイフン専用コード番号 U+2010 がマイナス専用コード番号 U+2212 が用意されているがこれらは使用していない。ただし、全テキストを通してマイナス記号は存在しないので、ハイフンとマイナス記号が混乱すること

はない。今後、テキストを追加した際にマイナス記号が含まれている場合には、両者を区別するために文字コードを変えることを検討する必要がある。

また、句読点として用いられる横棒(—)にはコード番号 U+2014(EMDASH)を用いている。ユニコードにはコード番号 U+2015(Horizontal Bar/QUOTATION DASH)が引用符用の横棒として用意されているが、これは用いていない。この理由は U+2014(EMDASH)は別の目的で使っていないので、EMDASH のための文字コードを用意している他の文字コードとの変換がたやすいということにある。EMDASH のための文字コードは cp1251、MacCyrillic には用意されているが、koi8-r、cp866 には用意されていない。koi8-r、cp866 では句読点用の横棒は HYPHEN_MINUS と合流してしまう。ユニコードの U+2015(Horizontal Bar/QUOTATION DASH)に対応する文字コードは cp1251、MacCyrillic、koi8-r、cp866 のどれにも用意されていない。

2.2. ヒダス方言

ヒダス方言の表記は基本的には標準語の表記法に従っている。ただし、ヒダス方言には標準語にはない音があるので、字素を組み合わせて新しい文字を作って表記することを試みている。

また、アバール語の標準語の正書法には問題があるので、その部分は修正してある。アバール語の標準語の正書法で最も大きな問題点は弱子音と強子音の区別である。アバール語の子音には弱子音と強子音の区別があり、アバール語の正書法では弱子音は子音字一字で表記するが、強子音は同じ子音字を二つ重ねて表記したり、弱子音と同様に子音字一字で表記したりする。たとえば、弱子音 κ に対する強子音を κκ で表記したり、κ で表記したりする。一般に、引用形(辞書に出てくる形)のレベルで弱子音と強子音の違いだけで意味が変わる場合には強子音を子音を重ねて表記する傾向がある。それ以外では、多くの場合、強子音を子音一つで表記することが多い。しかし、強子音を含むいくつかの語は対応する弱子音を持つ語が存在しなく、子音一つで表記しても意味を取り誤る可能性がないのに、強子音を子音を重ねて表記する。これは慣用としか言いようがない。また、強子音を一般的には子音を重ねて表記し、重ねないと違った意味に取られる語でも、ときどき強子音が子音一つで書かれていたりする。このような問題が起きないようにヒダス方言のテキストでは強子音は常に子音を二つ重ねて表記している。

電子化したヒダス方言のテキストでは、現在、便宜的に cp1251 を文字コードに用いている。この理由は、テキストの修正などのためにアバール語の母語話者とテキストのやりとりを行っていることにある。最終的には、utf-8 に変換を行う。

3. 電子化した資料

3.1. 標準語

電子化を行っている標準語の文献は次の 23 冊の本である。

- [A-1] Айтберов, Т. (1996) Цоралъул аваразул рагъазул тарих, МахIачкъала.
- [A-2] Аль-Къарахи, МухIамад-ТIагъир (1994) Дагъистаналъул хвалчабазул паркъи, МахIачкъала: Юпитер.
- [A-3] ГъалбацIов, ГъазимухIамад (1994) ГанчIал, МахIачкъала: Дагестанское книжное издательство.
- [A-4] ГъалбацIов, ГъазимухIамад (1994) Аварагзаби, МахIачкъала: Истина.
- [A-5] ГIалиев, Муслим (1993) Сардиль къвагъи, МахIачкъала: Юпитер.
- [A-6] Дадаев, Юсуп (1998) АхIул гохI - дир рекIел бухIи, МахIачкъала: Юпитер.
- [A-7] Даганов, ГIабдула (1997) ГIадамал - дир цIваби, МахIачкъала.
- [A-8] Жаватханов, Наби-Гулла (1993) Бадиса бадибе, МахIачкъала: Юпитер.
- [A-9] МуртазагIалиева, ПатIимат (1995) Кулакасул яс, МахIачкъала: Дагестанское книжное издательство.
- [A-10] МухIамадов, Муса (1991) Горо-цIер балелде цебе, МахIачкъала: Дагестанское книжное издательство.
- [A-11] МухIамадова, Майсарат (1996) Огъ, бихъинал, бихъинал, МахIачкъала: Юпитер.
- [A-12] МухIамадова, Сабигат (1992) Рокъи, МахIачкъала: Юпитер.
- [A-13] Расулов, ГIарип (1996) ГIадамалги рагIадалги, МахIачкъала: Дагестанское книжное издательство.
- [A-14] Расулов, КъурбангIаи (1997) БацIадисел, МахIачкъала.
- [A-15] Сурхаев, Мусалав (1990) Нух битIаги, МахIачкъала: Дагъучпедгиз.
- [A-16] Сурхаев, Мусалав (1994) Туснахъзда ГУЛАГалда, МахIачкъала: Юпитер.
- [A-17] Сурхаев, Мусалав (記載なし) Аварагасул халгIат, МахIачкъала: Юпитер.
- [A-18] ХIажиев, ХIусен (1995) Имам ХIамзат, МахIачкъала.
- [A-19] Шахтаманов, ГIумар-ХIажи (1994) Къарал гIор, МахIачкъала: Дагестанское книжное издательство.
- [B-1] ГIабдулаев, М. ГI., Г. И. Мадиева (1979) Авар мацI: 2 класс: ЦIали, грамматика, битIунхъвай ва калам цIебетIезаби, МахIачкъала: Дагъучпедгиз.
- [B-2] Меджидова, Ч. М. (1991) Авар адабияталъул чирахъ, МахIачкъала: Дагъучпедгиз.

[B-3] Мухтаров, С., А. Хамзатов, Ч. Меджидова (1991) Авар литература : 5 класс, МахIачкъала: Дагъучпедгиз.

[B-4] Раджабов, МухIамадIали Гъамбулатович (1990) Эркенаб заманалда цIалиялъе гIехъ : 4 класс, МахIачкъала: Дагъучпедгиз.

A-1 から A-19 は一人の著者の本であり、一冊の本が一作品からなっているものも、複数の作品が含まれているものもある。B-1 から B-4 は学校のアバール語に関する教科書や副読本であり、複数の著者の作品が含まれている。それぞれの文献は一つづつのファイルに納めてある。全体で 100 万語ほどである。テキスト A-12, A-9, A-10 の始まる部分をサンプルとして論文末にこの順番で載せている。

このほかに、雑誌 МагIарулай を 1 号分、新聞 ХIакъикъат を 1 号分電子化しているが、全く試験的な段階にある。

3.2. ヒダス方言

ヒダス方言のテキストは先にも述べたようにヒダス方言の話者に依頼して作文してもらったテキストであるが、一つの作品(話)を一つのファイルに納め、219 のファイルから成っている。全体で 10 万語ほどである。テキストのサンプルを論文末に載せている。

4. テキストへの注釈方法

テキストへの注釈の方法としては XML(Extensible Markup Language)を採用した。XML を採用した最大の理由は、XML がテキストを整形する方法のなかで最近最も使われているためである。一般に利用されているコーパスのなかでは ANC(American National Corpus)が XML を採用しているし、BNC(British National Corpus)も XML 版を追加しているところである³。また、いろいろなジャンルのテキストをマークアップする方法を標準化することを目指している TEI(Text Encoding Initiative)ももとの SGML (Standard Generalized Markup Language) 版に XML 版を追加している⁴。

XML 文書とは、簡単に言うと、開始タグ<x>と終了タグ</x>(x は任意の文字列)で囲まれた範囲に文字列あるいは別の開始タグと終了タグのペアが入っている文書である⁵。開始タグ<x>から終了タグ</x>までの範囲(開始タグと終了タグを含む)を要素(element)x と呼び、開始タグ<x>と終了タグ</x>に挟まれた部分(開始タグと終了タグは含まない)を要素 x の内容(content)と呼ぶ。開始タグには<x y="z">(x, y, z は任意の文字列)のように属性 y とその値 z のペアを複数個含めることができる。例を下に示す。

```
abc<element1 attribute1="value1">de</element1>fghi  
<element2 attribute2="value2" attribute3="value3">jk  
<element3>lm</element3>nop</element2>qr
```

この例は XML で整形されたテキストの断片である。改行は意味を持たないので、無視してよい。element1、element2、element3 が要素の名前(タグ名ともよく呼ばれる)、attribute1、attribute2、attribute3 が属性、value1、value2、value3 が属性の値である。要素 element1 は一つの属性とその値のペアを含んでいて、要素 element2 は二つの属性とその値のペアを含んでいるが、要素 element3 は属性とその値のペアを含んでいない。このように、要素は属性とその値のペアを含んでも、含まなくてもよい。element2 の開始タグと終了タグに挟まれた範囲に element3 の開始タグと終了タグに挟まれた範囲が含まれている。言い換えると、element2 の内容には要素 element3 が含まれている。

このように XML はテキストを階層的に整形することができる。XML はホームページを作成するのに使われる HTML(Hyper Text Markup Language)に似ているので HTML を知っている人にはなじみやすいと思われる。HTML は XML の前身であると言ってよい SGML に基づいていて、XML に基づく HTML の後継である XHTML(Extensible Hyper Text Markup Language)も最近使用され始めている。XML と HTML の大きな違いとしては HTML ではタグ名が限定されているのに対して、XML では任意のタグ名を使うことができること、HTML では終了タグを省略することができるのに対して、XML では終了タグを必ず書かなくてはならないことなどがあげられる。

XML は視覚的にわかりやすいことも重要である。研究を行いながら、テキストにいるような情報を注釈していくには、文書が視覚的にわかりやすいというのはとても重要なことである。

XML 文書の構造は DTD や XML スキーマなどを用いて規定することができる。筆者は現在 DTD を用いて文書の構造を規定している。

具体的には、次のような DTD である。

```
<!ELEMENT book (head, body)>
<!ELEMENT head (title, author, place, year, publisher)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT place (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT publisher (#PCDATA)>
<!ELEMENT body (part+|story+)>
<!ELEMENT part (title, story+)>
<!ELEMENT story (title, chapter+|(PARA|poem|LINE)+)>
<!ELEMENT chapter (title, section+|(PARA|poem|LINE)+)>
<!ELEMENT section (title, subsection+|(PARA|poem|LINE)+)>
<!ELEMENT subsection (title, (PARA|poem|LINE)+)>
<!ELEMENT PARA (SEN+)>
```

```
<!ELEMENT SEN ((w|PUN|PAGE)+)>
<!ELEMENT w (m+)>
<!ELEMENT m (#PCDATA)>
<!ELEMENT PUN (#PCDATA)>
<!ELEMENT PAGE (#PCDATA)>
<!ELEMENT poem (stanza+)>
<!ELEMENT stanza (verse+)>
<!ELEMENT verse ((w|PUN)*)>
<!ELEMENT LINE (#PCDATA)>
```

この DTD の意味する所は次節以下で述べることである。

5. 標準語のテキストの構造と要素名

5.1. 基本的な枠組み

XML 形式で注釈したアバール語の標準語のテキストは次のような構造をしている。
x は任意の字素、ハイフン、句読点、スペースを表す。

```
<?xml version="1.0" encoding="utf-8"?>
<book>
<head>
<title>xxxxx</title>
<author>xxxxx</author>
<place>xxxxx</place>
<year>xxxxx</year>
<publisher>xxxxx</publisher>
</head>
<body>
<part>
<title>xxxxx</title>
<story>
<title>xxxxx</title>
<chapter>
<title>xxxxx</title>
<section>
<title>xxxxx</title>
<subsection>
<title>xxxxx</title>
```

```

XXXXXXXXXX
</subsection>
<subsection>
<title>XXXXXX</title>
XXXXXXXXXX
</subsection>
</section>
</chapter>
</part>
</body>
</book>

```

文書の最も上のレベルは開始タグ<book>で始まり終了タグ</book>で終わる要素 book である。要素 book の次のレベルは要素 head と要素 body からなる。開始タグ <head>と終了タグ</head>の間に本についてのメタ情報が書かれていて、開始タグ <body>と終了タグ</body>の間が本の本文である。

要素 head は次の要素を並列的に含んでいる。

1	title	題名
2	author	著者名、編者名
3	place	出版地
4	year	出版年
5	publisher	出版社

表 3 文献の情報に関するタグ

<title>と</title>の間に本の題名が入る。<author>と</author>の間に単著である文献 A-1 から A-19 では著者の名前が、また、教科書等である文献 B-1 から B-4 ではいろいろな作家の作品をまとめた編者の名前が入る。そして、<place>と</place>の間には出版地、<year>と</year>の間には出版年、<publisher>と</publisher>の間には出版社の名前が入る。

要素 body より下のレベルは次のような階層になっている。

1	part	部	任意
2	story	作品(話)	必須
3	chapter	章	任意
4	section	節	任意
5	subsection	下位節	任意

表 4 作品の枠組みに関するタグ

<story>と</story>で挟まれた箇所が一つの作品(話)になる。要素 story は必須要素である。要素 part はいくつかの話をまとめるために使われ、<part>と</part>の間にはいくつかの作品(話)が含まれる。一部が短編小説、二部が中編小説というように作品(話)を分類するために使われている。要素 part は全ての文書で使われる必須要素ではなく、任意であり、多くの文書では要素 body の下に直接的に要素 story が来る。

要素 chapter が story の下に来る。この要素も任意であり、story の下に直接文章が始まることもある。chapter の下は section で、その下は subsection である。section および subsection も任意である。筆者が電子化した文献では subsection の下を設定する必要はなかった。subsection は section の存在を前提とし、section は chapter の存在を前提とする。すなわち、section がないのに subsection があることはないし、chapter がないのに section があることもない。

問題となるのは chapter の存在である。章と節の区別は難しいことも多いので、抽象的に考えると、chapter をなくして、section から始めて、以下は subsection、subsubsection と頭に sub を付けていった方がいいのかもしれないが、現段階では chapter から始めている。

いくつかの文献で *** で章、節、下位節を二つの部分に分けている場合がある。この場合、単に<LINE>***</LINE>として要素 LINE を付けただけで、chapter を二つの section あるいは section を二つの subsection に分けるというような方法も考えられるが、そのような方法は採用しなかった。要素 LINE は下で述べる要素 PARA と同じレベルになる。

ここまでは、章、節といった本、作品(話)の枠組みに関するタグ付け(注釈)である。このレベルまでのタグ付けは文献 A-1 から A-19、文献 B-1 から B-4 の全てに行っている。これより下は、段落、文、語というようなレベルになるが、このようなレベルへのタグ付けは一部の文献だけに実験的に行っている段階である。

5.2. 文章へ注釈

5.2.1. 要素 PARA: 段落

段落で問題になるのは、— で始まる会話の部分である。アパー語の本では、段

落の最初と — で始まる会話の最初は同じ幅で字下げされているので、区別することができない。したがって、— で始まる会話の始まりも段落の開始とみなしている。

実際にタグ<PARA>および</PARA>を入れている文献は A-8、A-9、A-11、A-12、A-15 だけであるが、それ以外の文献では改行記号が段落の切れ目を示しているので、簡単なプログラムを作って、残りの全文献にタグ<PARA>および</PARA>を自動的に入れることができる。

5.2.2. 要素 poem

いくつかの文献は文章の中に詩を含んでいるものがある。詩の部分はタグ<poem>と</poem>で囲んでいる。下に示すように、要素 poem は要素 PARA と同じレベルにしてあり、詩は段落の中に含まれていない。

```
<PARA>xxx</PARA>
<PARA>xxx</PARA>
<poem>xxx</poem>
<PARA>xxx</PARA>
```

poem 以下の階層は次のようになる。

1	poem	詩
2	stanza	連
3	verse	行

表 5 詩へのタグ

アバール語の詩は四行詩であり、四行が一連をなし、何連かが集まって一つの詩をなしている。二連からなる詩だと次のようになる。

```
<poem>
<stanza>
<verse>xxx</verse>
<verse>xxx</verse>
<verse>xxx</verse>
<verse>xxx</verse>
</stanza>
<stanza>
<verse>xxx</verse>
```

```
<verse>xxx</verse>
<verse>xxx</verse>
<verse>xxx</verse>
</stanza>
</poem>
```

詩を含む全ての文献で詩の部分に関しては要素 `verse` のレベルまではタグ入れを行っている。

5.2.3. 要素 SEN:文

要素 `PARA` の下のレベルは要素 `SEN` であり、要素 `SEN` は文を意味する。段落を文に分けるのがテキストを自動的に構造化、注釈する上で最も問題になる。また、言語間の差も大きいと考えられる。

文に分けるときに問題になるのは、「誰々が何々と言った」というように会話を含む部分で、何々の部分に複数の文が入っている場合に、どのように文を切るかということである。W を語とすると、次の(1-a)のような構造では(1-b)のように、また、(2-a)のような構造では(2-b)のように文をそれぞれ二つに切っている。

(1-a) — W W. W W — ян(「～と」) абуна(「言った」) W W.

(1-b) <SEN>— W W.</SEN>

<SEN>W W — ян абуна W W.</SEN>

(2-a) W W абуна(「言った」), — W W. W W — ян(「～と」).

(2-b) <SEN>W W абуна, — W W.</SEN>

<SEN>W W — ян.</SEN>

文をどのように切るべきかについては今後さらに検討を行う必要がある。

要素 `SEN` は文献は A-8、A-11、A-12 にだけ入れている。また、要素 `PARA` まで入れている文献 A-9、A-15 は一つの改行記号で文の切れ目を示すようにしてあり、文献 A-13、A-19 は二つの改行記号で段落の切れ目を示し、一つの改行記号で文の切れ目を示すようにしてあるので、これらの文献では簡単なプログラムを書いて、自動的にタグ `<SEN>` および `</SEN>` を入れることができる。

5.2.4. 要素 w、要素 PUN、要素 m

要素 `SEN` の下のレベルは要素 `w` および要素 `PUN` である。要素 `verse` 下のレベルも同様に要素 `w` および要素 `PUN` である。要素 `w` と要素 `PUN` は同じレベルにある。要素 `w` は字素とハイフンだけからなっていて、文の先頭、句読点、スペースではさまれ

たものである。

要素 PUN は句読点に付けてある。句読点は、... や?! のように複数個並ぶ場合がある。この場合、下の(1-b)、(2-b)のように複数個の句読点をまとめて要素 PUN を付けることも考えられるが、下の(1-a)、(2-a)のように一つ一つの句読点に要素 PUN を付けている。

(1-a) <PUN>.</PUN><PUN>.</PUN><PUN>.</PUN>

(1-b) <PUN>...</PUN>

(2-a) <PUN>?</PUN><PUN>!</PUN>

(2-b) <PUN>?!</PUN>

要素 w および要素 PUN の例として、次の(1)に一つ下のレベルのタグを入れると(2)のようになる。x は任意の字素である。

(1) <SEN>xxx xxx, xxx-xxx xxx.</SEN>

(2) <SEN><w>xxx</w> <w>xxx</w><PUN>,</PUN>

<w>xxx-xxx</w> <w>xxx</w><PUN>.</PUN></SEN>

要素 w の下のレベルは要素 m であり、要素 w は一つ以上の要素 m から成っている。要素 w が二つ以上の要素 m から成っている場合、一番目の要素 m が語彙的な意味を持った部分で、二番目からの要素 m は ги「も」、цин「さえ」、強調を表す го などの語彙的な意味を持たない付属形式である。以後、語彙的な意味を持つ自立形式を自立語と呼び、ги「も」、цин「さえ」、強調を表す го などを付属語と呼ぶことにする。要素 w のレベルを語と呼ぶべきか、要素 m のレベルを語と呼ぶべきかはいろいろと議論があると思われるが、ここでは問題にしない。また、付属語として自立語から切り離す部分についてもある程度恣意的であり、格を表す変化語尾や時制を表す変化語尾などは自立語の一部として切り離していない。こうした問題は今後十分に検討する必要がある。

要素 w および要素 PUN をプログラムを書いてタグを付けるのは簡単であるが、要素 w を要素 m にプログラムを書いて分けるにはいくつかの関門がある。たとえば、берцин は形容詞(短語尾形)あるいは副詞 берцин「美しい」、「美しく」である場合と名詞 бер「目」と付属語 цин「さえ」の連続である場合があるので、下の(1)、(2)のように二つの区切り方があり、これは文章の内容を理解しないと区別することができない。

(1) <w><m>берцин</m></w>

(2) <w><m>бер</m><m>цин</m></w>

5.2.5. 要素 PAGE

要素 PAGE はページ番号を表す。要素 PAGE は要素 w および要素 PUN と同じレベルに設定している。ページ内の最後の要素 w が要素 PUN の後に PAGE タグを付けてページ番号を示している。

6. 要素 m に対する属性

要素 m には品詞に関する情報をはじめいくつかの属性を付加している。ここでは、そうした属性について説明する。

6.1. 品詞

品詞は属性 pos を用いて示している。語を自立語と付属語に分けて説明する。付属語とは前にも述べたように ги「～も」などのように自立語の後ろに付加して表記される語である。自立語の品詞は次の表のように分けている。

1	v	動詞
2	n	名詞
3	j	長語尾の形容詞
4	js	短語尾の形容詞
5	h	кІваричІю
6	b	副詞
7	p	後置詞
8	c	接続詞
9	int	感嘆詞
10	pn	代名詞
11	refl	再帰代名詞
12	rec	相互代名詞
13	num	数詞
14	week	曜日の名前
15	eth	民族などの名前
16	nv	動詞から派生した名詞
17	nj	名詞として使われている長語尾の形容詞
18	bj	副詞的に用いられている長語尾形容詞
19	jn	名詞の属格に形容詞の長語尾の接辞を付けた形容詞
20	jpn	代名詞の属格に形容詞の長語尾の接辞を付けた形容詞

表 6 自立語における品詞

この品詞分類は筆者の研究上の関心によるものが大きい。共同利用する場合には、いくつかの品詞を合流させ、別の属性を追加する必要があると考えられるが、属性を複数個並べると視覚的にわかりにくくなるので、現段階では便宜的に品詞を細分している。

形容詞は長語尾形容詞 *j* と短語尾形容詞 *js* に分けている。長語尾形容詞は、多くの場合、短語尾形容詞に *a=AM/я=AM* をつけることによってできる。*=AM* および *=AM=*、*AM=* は一致標識を意味する。アバール語の一致は男性(*в*)、女性(*й*)、非人間(*б*)、複数(*р*, *л*)の四分類から成る。長語尾形容詞は名詞を修飾する用法と存在動詞 *AM=укИне* 「ある、いる」とともに用いて述語としての用法を持っている。短語尾形容詞は名詞を修飾する用法を持たず、存在動詞 *AM=укИне* とともに用いて述語としての用法と *гъа=AM=изе* 「する、作る」とともに用いて「～を～にする」という意味を表す用法とを持っている。

品詞 *h* には *кИваричЮ* 「必要ない」だけが入る。この語は変化のパターンから動詞とも形容詞と分類できず、一つの語彙項目で一つのグループをなしている。

品詞 *eth* には *гИрус* 「ロシア」、*немец* 「ドイツ」、*гАраб* 「アラブ」、*турк* 「トルコ」、*гуржи* 「グルジア」、*чачан* 「チエチエン」、*лъарагI* 「クムイク」、*дарги* 「ダルギ」、*лезги* 「レズギ」などが含まれる。これらは、そのままの形で下の(1-a)、(2-a)のように *мацI* 「言語」、*гЕдму* 「文化」などいろいろな名詞を修飾することができる。(1-a)、(2-a)はそれぞれ(1-b)、(2-b)のようなタグが付けてある。

(1-a) *гИрус мацI*

(1-b) `<w><m pos="eth">гИрус</m></w> <w><m pos="n">мацI</m></w>`

(2-a) *гАраб гЕдму*

(2-b) `<w><m pos="eth">гАраб</m></w> <w><m pos="n">гЕдму</m></w>`

品詞 *nv* は動詞から派生した名詞である。たとえば、動詞 *хIалтИзе* 「働く」から派生した *хIалтИ* 「仕事」などである。

品詞 *nj* は名詞として使われている形容詞を表す。アバール語ではいろいろな形容詞を名詞として使うことができる。たとえば、形容詞 *берцина=AM* 「美しい」を「美しい人」、「美しいもの」のように名詞として使うことができる。アバール語の形容詞は名詞を修飾している場合には格変化をしないが、名詞として使われている形容詞は格変化をする。*берцинай ясаль* は形容詞 *берцина=AM* 「美しい」の女性形 *берцинай* に *яс* 「女の子」の能格形 *ясаль* が続いているが、形容詞は修飾している名詞の格を引き継いでいない。形容詞 *берцина=AM* 「美しい」を「美しい女性」という意味で用いると *берцинай* というように能格の形を取れるようになる。

品詞 *bj* は副詞的に用いられている形容詞である。たとえば、形容詞 *дагъа=AM* 「少

しの」の絶対格形 дагъа=AM あるいは能格形 дагъаль を副詞的に使うことができる。

品詞 jn は名詞の属格に形容詞の長語尾の接辞 a=AM を付けた形容詞である。たとえば、рагъ「戦争」の属格 рагъул に a=AM がついて рагъула=AM ができる。属格 рагъул とそれに a=AM がついた рагъула=AM の基本的な意味は同じであるが、分布に違いがある。

品詞 jpn は人称代名詞の属格に形容詞の長語尾の接辞 a=AM を付けた形容詞である。たとえば、дун「私」の属格 дир に a=AM がついて дира=AM ができる。属格 дир とそれに a=AM がついた дира=AM の基本的な意味も同じであるが、分布に違いがある。

付属語の品詞分類は次の通りである。

1	etc
2	quot
3	z
4	start
5	end
6	time
7	compar

表 7 付属語における品詞

付属語の品詞分類も筆者の関心を強く反映している恣意的なもの、そして便宜的なものになっている。2-7 の品詞に分類した付属語の余りを 1 の品詞 etc に分類している。品詞 etc には ги「も」、цин「さえ」、強調を表す го をはじめ 20 を超える付属語が含まれる。

品詞 quot には会話を引用する標識であり、日本語の「～と言った」の「と」に対応する ан/ян, (й)ин, (й)илан が含まれる。

品詞 z は疑問文の従属節を閉める標識である али/яли だけからなる。

品詞 start は алдаса/ялдаса「から」だけである。時間を表す副詞などについて開始時間を表す。たとえば、副詞 гъанже「今」に ялдаса「から」が付いて гъанжеялдаса「今から」ができる。

品詞 end は алде/ялде「まで」だけである。時間を表す副詞などについて終了時間を示す。たとえば、副詞 метер「明日」に алде「まで」が付いて метералде「明日まで」ができる。

品詞 time は алда/ялда だけである。数字について「～時に」を表す。たとえば、次の (1-a) は carIar「時間<絶対格>」と цо「1」に ялда が付いた цоялда からなっている。(1-a) にタグを付けると (1-b) のようになる。

(1-a) carIar цоялда「一時に」

(1-b) <w><m pos="n">carIar</m></w> <w><m pos="num">цо</m><m pos="time">ялда</m></w>

品詞 compar は алдаса/ялдаса「より」だけである。副詞や名詞の能格、与格などの変化形などについて比較の対象を表す。たとえば、副詞 себе「前」に ялдаса「より」が付いて себеялдаса「前より」ができる。

文献 A-8, A-11, A-12 では、全ての要素 m に品詞に関する属性を入れてある。ただし、現在行っている作業を効率よくするため、次の(1-a)を(1-b)に(2-a)を(2-b)にというように、全ての要素 m を品詞属性の値に変えた状態にしてある。全ての要素 m に品詞属性を入れると視覚的にわかりにくくなるためである。

(1-a)<m pos="n">xxx</m>

(1-b)<n>xxx</n>

(2-a)<m pos="v">xxx</m>

(2-b)<v>xxx</v>

もちろん、最終的には、元の状態へもどすことになる。

6.2. 誤植などの誤りへの対処

誤植などの誤りがある場合には、タグ<m>に correct 属性と org 属性を付けて示している。たとえば、тПоцebeseb「最初の」とあるべきところに тПецebeseb とある場合、下のような注釈がつけてある。

<m correct="y" org="тПецebeseb@">тПоцebeseb</m>

correct 属性の値は常に y であり、org 属性の値は誤りの後に@を付けたものである。@は XML 形式を考慮しない形で単純ながら高速の検索をするためにつけているメタ記号である。

句読点の誤りは注釈なしに訂正してある。たとえば、xxx, xxx とあるべきところに xxx ,xxx のようにコンマとスペースが反対になっているような誤りがあるが、このような誤りは訂正しているだけで、訂正したことを示す注釈は入れていない。また、xxx-xxx というハイフン入りの語の代わりに xxx — xxx というように、ハイフンがスペース、句読点用の横棒、スペースという連続になっている場合があるが、この場合も単に誤りを訂正しただけで、訂正したことを示す注釈を入れていない。

誤りではないが、行末の単語を割るために使われているハイフンは注釈なしで消去している。これについては、注釈の必要性を検討している最中である。

6.3. 引用形

一つの語彙項目が格や時制などによっていくつかの変化形として実現するわけであるが、一つの語(語形)が複数個の語彙項目の変化形であることがある。この場合、属性 base を用いて引用形(辞書で現れる形)を示すことを目指している。引用形は名詞では単数・絶対格形であり、動詞では不定形である。引用形は次のようにタグ<m>に base 属性を付けて示している。

```
<m base="xxx@">xxx</m>
```

base 属性の値は引用形の後に@を付けたものである。@は前に述べた org 属性の場合と同様に XML 形式を考慮しない形で単純ながら高速の検索をするためにつけているメタ記号である。

たとえば、動詞 AM=ахъизе「はずす」と AM=ахъине「立つ」の過去形はともに бахъана となるので次の(1)と(2)のようになる可能性がある。

```
(1) <m base="AM=ахъизе@">бахъана</m>
```

```
(2) <m base="AM=ахъине@">бахъана</m>
```

現段階で属性 base を付けるのに取り組んでいるのは八組の動詞と一組の名詞だけである。

次の表の動詞の組は不定形では違う語形であるが、一部の変化形で同じ語形になるので区別が必要である。この 8 組の区別を文献 A-8, A-11, A-12 および文献 A-9, A-13, A-15, A-19 において完全ではないが行っている。

1	AM=ахъизе	AM=ахъине
2	AM=елъизе	AM=елъине
3	AM=ильлъизе	AM=ильлъине
4	AM=иччизе	AM=иччазе
5	къине	къазе
6	цъезе	цъузе
7	ккезе	кквезе
8	гъа=AM=изе	гъаризе

表 8 区別する動詞の組

アバール語の動詞は語幹と時制などを表す変化語尾からなる。不定形の変化語尾は動詞によって-изе, -ине, -езе, -ене, -зе などがあるが、不定形の変化語尾によって他の変化語尾の形が違ってくる。

AM=ахъизе「はずす」とAM=ахъине「立つ」は語幹が同じであるが不定詞の変化語尾が違っている。過去形はともにAM=ахъанаとなり、それ以外もいくつかの変化形で同じ形になる。AM=елъизе「笑う」とAM=елъине「塗る」も同様に過去形AM=елъанаの他いくつかの変化形で同じ形になる。AM=ильльизе「小便する」とAM=ильльине「歩く」も同様に過去形AM=ильльанаの他いくつかの変化形で同じ形になる。

AM=иччизе「ぬれる」とAM=иччазе「放つ」は過去形がともにAM=иччанаになる他いくつかの変化形で同じ形になる。

къине「枯れる」とкъазе「閉じる」も過去形がともにкъанаになる他いくつかの変化形で同じ形になる。

цезе「満たす」とцүзе「濾す」も過去形がともにцүнаになる他いくつかの変化形で同じ形になる。

ккезе「起こる」とкквезе「つかむ」は現在形がともにкколаになり、その他のいくつかの変化形でも同じ形となる。кквезеのように子音の後にвがついていると前の子音が円唇であることを表す。кквезеのように語幹の最後の子音が円唇である語は変化語尾の最初の音がy[u]やo[o]などの円唇母音であるとккезеのように語幹の最後の子音が円唇でない語の変化形と区別がつかなくなる。

гъа=AM=изе「する、作る」は一致標識を含んでいるが、一致標識が非人間である場合不定形がгъаризеとなりгъаризе「頼む」の不定形と同じになる。他の多くの変化形でも同じ語形になる。

これ以外にも不定形以外の変化形で語形が同じになる動詞の組は少数ながら存在する。

名詞で区別を行っているのは名詞жал「たてがみ」と名詞жо「もの」の区別である。жалという語形は名詞жал「たてがみ」の単数・絶対格形である場合と名詞жо「もの」の複数・絶対格形である場合がある。また、再帰代名詞жи=AM「自分」の複数・絶対格形である場合もある。不定形で違う形を取る名詞で一部の変化形で同じ形になるものもこの他に多少ながら存在する。

ここまでがbase属性を付加することに取り組んでいる語彙項目であり、以後はこれからの課題である。

同じ語の違う変化形が同じ形になることがある。たとえば、малалという語形は名詞мал「足」の単数・属格形である場合と複数・絶対格形がある。同様に、түлалという語形は名詞түл「肝臓」の単数・属格形である場合と複数・絶対格形がある。単数・属格形と複数・絶対格形が同じ形である名詞はそれほどないが、最低でも20語ぐらいはある。

変化形の一部が同じ形になる語彙項目の組は品詞が違う場合もある。たとえば、бакIалという語形は名詞 бакI「場所」の複数・絶対格形である場合と形容詞 бакIа=AM「重い」の複数形である場合とがある。

また、лъаралという語形は名詞 лъар「川」の単数・属格形である場合と名詞 лъар「川」の複数・絶対格形である場合と動詞 лъазе「知っている」の形容詞的分詞・過去・複数形である場合がある。

base 属性を完全に付与するためには、同じ語形を持つ語彙項目を網羅的に調査する必要性を感じている。

6.4. その他の属性

上で述べた以外に研究の関心からいろいろな属性を付与しているが、どれも試験的なものであるので、ここでの説明は行わないことにする。

7. ヒダス方言のテキストの構造

ヒダス方言のテキストは現在も母語話者とのやりとりを行い、修正、改良を行っている段階にある。これは、筆者がヒダス方言について研究を始めたばかりで、全体的によく理解できていないことと関係する。

こうした段階にあるので、テキストへのタグに関してはタイトルなどにタグを入れているくらいである。ただし、段落の切れ目は改行記号二つで示している。

テキストにはテキストの作成者などの情報は入れてなく、別のファイルにそれぞれの作品(話)の名前、執筆者名、執筆年、ファイル名をリストアップしている。

ある段階で標準語のテキストと同じ構造でタグを入れ、XML 文書にする。

8. おわりに

以上述べてきたような形で、アバール語の標準語とヒダス方言のテキストの電子化を行ってきた。最初にテキストの電子化を始めたとき、また、段落より下のレベルのタグを入れ始め、品詞などの属性を入れ始めた時に考えていたよりも遙かに時間がかかる作業であった。今後、さらに徹底したタグ入れを行うとともにいろいろな問題点に取り組んでいきたい。また、標準語のテキストは電子化する段階で写し誤った箇所がまだたくさん残っているので、タグを入れながら、訂正を行っていく。

注

1. 本稿ではアバール語の文法について説明しないので、アバール語の標準語に興味のある方には Alekseev & Ataev(1997)、Madieva(1980)が参考になる。残念ながら、日本語、英語で読めるものは存在しない。また、アバール語の諸方言について

- は Mikailov(1959)が参考になる。
2. ユニコード(Unicode)についての公式の情報はサイト[1]で得ることができる。
 3. ANC と BNC についての公式の情報はそれぞれサイト[2]とサイト[3]で得ることができる。
 4. TEI についての公式の情報はサイト[4]で得ることができる。TEI の紹介は Ide & Jean (ed.)(1995)にある。
 5. XML についての公式の情報はサイト[5]で得ることができる。また、XML の基礎については中山、奥井(編著)(2001)などで学ぶことができる。

参考文献

- Alekseev, M. E. & B. M. (1997) *Ataev Avarskij Jazyk*, Moskva: Academia.
- Ide, Nancy & Jean Véronis (ed.), (1995) *Text Encoding Initiative: Background and Context*, Dordrecht: Kluwer Academic Publishers.
- Madieva, G. I. (1980) *Morfologija Avarskogo Literaturnogo Jazyka*, Maxachkala: Daguchpedgiz.
- Mikailov, Sh. I. (1959) *Ocherki Avarskoj Dialektologii*, Moskva, Leningrad: Izdatel'stvo Akademii Nauk SSSR.
- 中山幹敏、奥井康弘(編著) (2001) 『改訂版標準 XML 完全解説(上)』、技術評論社。

参照サイト(2006年3月25日現在)

- [1] <http://www.unicode.org/>
- [2] <http://americannationalcorpus.org/>
- [3] <http://www.natcorp.ox.ac.uk/>
- [4] <http://www.tei-c.org/>
- [5] <http://www.w3.org/TR/REC-xml>

標準語のテキストのサンプル

以下のテキストで¶は改行記号を表す。

例 1 : A-12

```
<?xml version="1.0" encoding="utf-8"?>¶
<book>¶
<head>¶
<title>Рокъи</title>¶
<author>Сабигат МухІамадова</author>¶
<place>МахІачкъала</place>¶
<year>1992</year>¶
<publisher>Юпитер</publisher>¶
</head>¶
<body>¶
<story>¶
<title>АХИРИСЕБ ДАНДЧИВАЙ</title>¶
<PARA>¶
<SEN>¶
<w><n>Заман</n></w> <w><v>щвелалде</v><etc>го</etc></w>
<w><v>херлъарай</v></w>
<w><n>ХІабибат</n></w><PUN>,</PUN><w><j>бакІаб</j></w>
<w><n>сумка</n><etc>ги</etc></w>
<w><v>босун</v></w><PUN>,</PUN><w><n>больницаялъул</n></w>
<w><n>кІалтІе</n></w> <w><v>щвана</v></w><PUN>.</PUN>¶
</SEN>¶
<SEN>¶
<w><b>Гъениб</b></w> <w><num>цо</num></w> <w><n>гъутІбузда</n></w>
<w><p>гъоркъ</p></w> <w><j>халатаб</j></w> <w><n>бакІида</n></w>
<w><b>гІодой</b></w> <w><v>чІана</v></w><PUN>.</PUN>¶
</SEN>¶
<SEN>¶
<w><j>ГІемерал</j></w> <w><n>пикраби</n></w> <w><v>рукІана</v></w>
<w><pn>гъальул</pn></w> <w><n>ботІроль</n></w> <w><v>
meaning="n">хенелел</v></w><PUN>.</PUN>¶
</SEN>¶
<SEN>¶
```

<PUN>«</PUN><w>ГъадигІан</etc>го</etc></w>
 <w><j>захІматаб</j></w> <w><n>иш</n></etc>ищ</etc></etc>ха</etc></w>
 <w><v>унтарав</v></w> <w><n>чиясде</n></w> <w><v
 check="y">яккизе</v></w>
 <w><v>ин</v></w><PUN>?</PUN><PUN>»</PUN><w><v>абун</v></w>
 <w><refl>жинца</refl></etc>го</etc></w>
 <w><refl>жиндие</refl></etc>го</etc></w> <w><n>суал</n></w>
 <w><v>къолеб</v></w> <w><v>букІана</v></w>
 <w><pn>гъель</pn></w><PUN>...</PUN>¶
 </SEN>¶
 </PARA>¶
 <PARA>¶
 <SEN>¶
 <w><n>РакІалде</n></w> <w><v>щвана</v></w> <w><j>рикІадесеб</j></w>
 <w><n>инсул</n></w> <w><n>росу</n></w><PUN>.</PUN>¶
 </SEN>¶
 <SEN>¶
 <w>Кида</etc>дай</etc></w> <w><n>ХІабиб</n></w>
 <w><n>рокъове</n></w> <w><v>вачІилеван</v></w> <w><n>нухда</n></w>
 <w><n>берал</n></etc>ги</etc></w>
 <w><v>лъун</v></w><PUN>,</PUN><w><n>рагъида</n></w>
 <w>гІодой</w> <w><v>чІун</v></w> <w><v>йигей</v></w>
 <w><refl>жий</refl></etc>го</etc></w><PUN>.</PUN>¶
 </SEN>¶
 <SEN>¶
 <w><refl>Жиб</refl></etc>го</etc></w> <w><n>лъимадул</n></w>
 <w></etc>ГІадин</etc></w> <w><v>хІакъдулеб</v></w>
 <w><v>букІараб</v></w> <w><n>ракІ</n></w><PUN>.</PUN>¶
 </SEN>¶
 </PARA>¶
 <PARA>¶
 <SEN>¶
 <w><j>ТюкІаб</j></w> <w><n>гІумрюяль</n></w>
 <w><n>жо</n></etc>го</etc></w> <w><v>букІинчІеб</v></w>
 <w></etc>ГІадин</etc></w><PUN>,</PUN><w><pn>гъеб</pn></w>
 <w><v>щолаан</v></w> <w><n>ракІалде</n></w><PUN>.</PUN>¶
 </SEN>¶

例 2 : A-9

```
<?xml version="1.0" encoding="utf-8"?>¶
<book>¶
<head>¶
<title>КУЛАКАСУЛ ЯС</title>¶
<author>ПатИмат МуртазаГалиева</author>¶
<place>МахАчкъала</place>¶
<year>1995</year>¶
<publisher>Дагестанское книжное издательство</publisher>¶
</head>¶
<body>¶
<part>¶
<title>КЪИСАБИ</title>¶
<story>¶
<title>КУЛАКАСУЛ ЯС</title>¶
<chapter>¶
<title>СУАЛ</title>¶
<PARA>¶
Рогъалильго росулъа къватИреги <m pos="v" base="АМ=ахъине@">рахъун</m>,
магъилъа эхере унел рукІана Гумарги ПатІнаги.¶
Росги лъадиги гІедеГун рукІана колхозаль жидеего бикъун къураб хурул
бутІаялде, гІемераб букІиндалха гъениб хІалтІи.¶
ГІадибалил тухумалъул букІараб хуралде гІагарлъараб мехалда ПатІна лъалхъана,
цо щибалиго кІвар бугеб жо ракІалде щвараб гІадин.¶
</PARA>¶
<PARA>¶
– Гъа, кие балагъарай?¶
Щиб бихъулеб бугеб? — ан цІехана Гумарица.¶
</PARA>¶
<PARA>¶
– Гумар, ноль-къаси гъагъаб гъобоялъул гІохдасан беролеб букІараб кІкІуй
бихъизе лъикІан дуда...¶
</PARA>¶
<PARA>¶
– Цо-цо гІадалал <m pos="n" base="жо">жал</m><m pos="etc">ги</m> рицуна
дуца.¶
```

НекІого рехун тараб гьобоги гьединабго хурги!¶

Щив гІабдал унев гьоба цІа бакизе?¶

Гьа-гьа-гьа!¶

Макьиль бихьун батила!¶

ЧІанда цІалкІичІого дурго нухда <m pos="v"

base="АМ=ильльїне@">йильльга</m>!¶

</PARA>¶

<PARA>¶

– Дир махсаро гуро!¶

Макьиль бихьичІо, хІакълъун кІкІуй беролѐб букІана.¶

Тамаша-гІаламат, гьобо гьададинго буго.¶

Дидани <m pos="v" base="ккѐзе@">ккун</m> букІана гьаниб цІа <m pos="v"

base="ккѐзе@">ккун</m> батилилан.¶

<m pos="v" base="АМ=ильльїне@">Рильльгая</m> цо балагьїзе...¶

</PARA>¶

<PARA>¶

Заман гьѐчІѐб мехалда кьваригІѐл гьѐчІѐб иш батарай лъадуде

реклѐльготултудилев вукІаниги, гьѐлда хадув кьокъана ГІумар.¶

Гьабирокъобе нух кьан батидал гьѐвги тамашалъана.¶

ГІумарие захІмат букІинчІо гьенир лъурал ганчІалги басралъарал хьарщалги
нахъе рехїзе.¶

Нухги гьабун жанїве лъугъарав рос тІун кьѐргун нахъе-нахъе кьан вачІараб
мехалда хІинкъана ПатІїнаги.¶

</PARA>¶

<PARA>¶

– Доя... доя чІужугІадан йїго... хварай лъаларо, чІварай лъаларо, — ян абуна гьѐс,
бабадулаго мацІ гун.¶

</PARA>¶

<PARA>¶

КигІан мухІканго балагъаниги, я ГІумарида я ПатІїнада лъачІо хварай кІодо
щїяли.¶

КигІан ургъаниги бичІчІичІо щай гьѐй рехун тараб гьабїхъе ккараяли.¶

ЧІванищ, унтунищ гьѐй хун йїгѐяли.¶

</PARA>¶

例 3 : A-10

```
<?xml version="1.0" encoding="utf-8"?>¶  
<book>¶  
<head>¶  
<title>Горо-цлер балелде себе</title>¶  
<author>Муса МухАмадов</author>¶  
<place>МахАчкъала</place>¶  
<year>1991</year>¶  
<publisher>Дагестанское книжное издательство</publisher>¶  
</head>¶  
<body>¶  
<chapter>¶  
<title>БАЙБИХЪИ</title>¶  
<section>¶  
<title>1</title>¶
```

ПатИдаб горда нахъа, гИсинал гъакал гъоркъ рараб бакИда хъувухъун гЮдов чЮн вукАна Хъзаами. ППолохъанаб мехаль гАнабазул тЮгъздегицин рагАдги рехизабун, хуч-хучун рукАрал халатал тЕлхалги хЮлун, тАмах гъараб гъотЮда хутАрал цЮрорал пихъил гарал гАдин, гЮнун ругел гъесул чАхИиял берал чИваркъун рагъдухъе ралагъун ругоан. Цодагъаб цебеги теретлъун хъвагЮлел рукАрал гАзул гарал гъанже, гъава къваранагАн гЕдерлъун, цИвабазул суратги лъугъун божигун роржунел рукАна. Ракъалде щвейдал щ у лаго цоцаль хурхун чЮлел ругоан гъел; жеги-харги гЮдосан чИвалеб хинлъиялъ жалго хваниги, дол хадур рачЮнел чАго хутИзе гЮло.¶

— Цо гАкъилаб кАлалъ гАдада абун гъечЮ, цебесеб гЕл хадур гЮлезе гЮмруялде лъураб къольун букЮнилан, — угъун биччана Хъзаамица, мугъзакъ лъураб хъупил къандалъоги битА-бишизабулаго. Гъадал гАзул хЮлабазул хАкъикъаталъ хИкмалъизавун вукАна гъев. Гъеб хИкматалъ рижизарурал пикрабаз заманаялъ къинабуна ахАнжеги гъасул рекЕл рикКлада хутАраб бахАрлъиялда хадуб бугеб угъди. — ГАдамазулалци щай букЮнареб. гИсинал рукЧАголъаби — цЮнцАрабазулгицин буго гъеб багъадураб хасият. Жидерго нухда бахине кЮлареб «кКАл», «хъитИ» яги кИбекараб гъотЮл гАркъелалъ лъугъинабураб квалквал себе батарабго, цоцалье рутИбиги риччан регун къо гъабун чЮла гъел, цойгидаб жидерго гъалмагълъи тАсан ун бахъинегАн. КигАн кЮдияб бакИлъи тАде кканиги тЕзе толаро гъез гъеб къо.¶

— Гъал тАдагъал, цохЮ пуни гъородахъ унел гАзул хЮлабазулги хАлкЮлъи

бихьуларич? — илан хьвагІезабуна Хьзаамица надалдеГанисан ГІебаб,
ГІоноцІиса бегІерал, гьваншида хутІарал халатал расалги кІиго-лъабго свери
гьабун ГІаде рачун, ГІубан хІулун, гьечІеблъи бахчизе хІаракат бахьулеб бетІер. —
Балагье дудаго гьозухь! Цоцалье гьегжи биччан гьат гьабун регулел руго,
гьородаги гьагьаб ракъалдасан чІвалеб хинлъиялдаги дандечІезе къуват гьеб
цолъиялъулъ букІин бичІчулеб ГІадин. Доле чІухІараб парпаригун гьаваялдасан
роржун рачІун батІаго [[3]] рещтІунел цо-цо ГІазул гарал, сапнал полпал ГІадин,
рихьунгн ун, хьахІаб куц-мохьги ГІагІун, лъадал ГІирабаздеги руссун холел
руго...¶

Циндаго бахьараб нуцІил цІурмил гьаракъаль гьоркъор къотІизаруна Хьзамил
пикраби. ГьебсаГІатго «Хьартилги» бахьана хІапи. НуцІихье бортилелде, цин
ГІепизеги гьабун Хьзаамихь балагьана, рукъальул букІнида жиндирго бакІалда,
басрияб ГІансаялда ГІад кІусун букІараб гьеб. «Дур гьаниб щибго хІажат
гьечІилан хІапдон нахьейищ къотІилев гьобол яги къабулищ гьавилев? Дун дур
буюрухьалъухь балагьун буго», — ян гьикъулеб букІана залимаб черхалъул,
чІегІераб ГІомалъул, гьаб рукъоб къоял гурел, сонал рараб, гьединлъида л
гьалъул хважаинзабазул хасият лъикІ лъалеб гьойдуца Хьзаамида.¶

— Огь! Подхалим! Загьидат рукъой йигеб мехаль хІелхІедулароан мун букІараб
дие гьедин! — илан, гьелъул рагьа-ракариялъулъ жиндирго ГІамалалъул раГІадги
бихьун, рокьукъго валагьана Хьзаами «Хьартихь». — ЛъикІавги квешавги чи
рагьдухье щолаго ватІа гьавизе гьунар букІана дур себе. Гьанже ГІехІун буго
дурги цІодорлъи, — ян, ГІемерисеб мехаль жиндиего тахлъун бугеб, тамахаб
бакІги гиризабулаго жанахІалъувехун нух босана гьес. «ЩивГІагидай
ГІага-божаразцин рехун тарав дихье вачІунев вуго? Цосинав вагьа-вакаризе гьунар
гьечІев дунгоги, дир къоялде ккечІониги, тату хвараб гьойги гурони рукъор
гьечІеллъиги лъан вачІунев цІогьорцин вугодай?»¶

Хьзамил лъади туристазул къукъагун цадахь Грециялде аралдаса анцІабго къо
букІана. «Ралъадалги рахун, гьаваги къотІун, гьединачал халатал сапаразде яхьунеб,
къуват гьечІин дур», — илан абиялъул чІечІо. «БахІараб мехаль мунги арав
батІиял улкабазде. Диеги бокьун буго дунялги, бихьизе, дагьабниги гьогьенги
чІвазе», — ян: абуна гьель.¶

— Гьогьен чІвазе бокьани, дурго росулъе КІуситІе а, яги дир лъимерлъи араб
КьуртІаколосе къокъа. Эбел-инсул хабахухьеги щевезе, себе кІочон тараб маГІарул
ракъальул махІги сунтІизе, рекІелъе сабруги рещтІунеб, черхги чІаголъулеб куц
бихьизе, — ян абуна хадуб Хьзаамица.¶

— Эбел-инсул хабзалахьеяли хадуй щелин, цин Одиссейил улка бихьизе бокьун
буго, — ян, йиго гьеб мехальги жаваб къолей.¶

ヒダス方言のテキストのサンプル

<title>Билкьисги кьурачги.</ title >

¶

Со Кьурачмамо цIар буго чияссе ячумо яГIана гьессе йокьумо, эбел-энссол щуго-анлъльго васаза гьокьо сохIо йиге Билкьисмамо цIар буге яс. Гьелъльеги бокьумо бугIумо бугоха гьессе эне. Билкьисги ячумо Кьурач шагьаралълье ана, гьени хIалтIинаги ваГIана. Гьел сосазе кьиматаа, цIакъ лъльикI ГIумро гьуна раГIана. Гьедин данде руссумо гьезе кIиго васги гьувуна. Со заманаяассан бугI аб хасият бегьина гьесул чорхольлъе, мукилдале байбихьина. Садаха кьутIире рахумо гьудулзабигимо гьев гвердена вугонидин кьIаде йитIумо гьез`э Билкьис балале йегьина йигIумо гьео. Рукье щварамалъльа хабар гьубуна бугIумо буго. Со пуланаб кьоялъ Кьурачил эбел ГIунайзат егьина росулъльа шагьаралълье. Эбелги ячумо Кьурач ана зияраталъль.¶

¶

Гьел нах руссина заман Билкьисида лъана баГIанаро. Билкьис Кьурачги эмо хIалкъамо теого гьелъухе регьина мадугьалзаби. Гьезе Билкьисид гьоболлъи гьубуна, ГIодо чIемо хабар-кIалаа раГIана. Гьел киналго гьалбал наха мехх щвемо бугомамо рукьи-рукьире ана.¶

¶

Билкьисид мадугьал ГIумар гьени чIчIеле гьуна лъималаз хIвеле гьубураб магнитафон кьачIале мамо. Гьеб кьачIана гьев ваГIараго нуцIцIада кIутIина. Билкьис йиххумо хутIина, гьав щодай нахамалъа ваино гьоболмамо. НуцIида наху-нахуги кIутIкIутIина. Билкьис хIинкьараго нуцIцIа ричIле ана. НуцIцIа ричIараасаго жане вегьинаха цIакъ ссин бохумо Кьурач. Кьурач вегьараасаго эхIделе байбихьина: «Що гьани вуго? Щогимо мун йиге?». Билкьис: «ЭхIдего! Що гьани вугIнахо, нилъер мадугьал вугоха!». Мадугьал дун рукьи гьеамалъа гьане щиле вегьумо вугомамоги абумо Кьурачид ГIумариэ нус бахина. ХIинкъумо Билкьисид гьаркьал ращана. Гьел гьаркьал раГIумо мадугьаллъи бегьина, ва Кьурачиха нус нахе босина. Билкьисид Кьурачида бичIчIле гьубуна баГIана ГIумар магнитафон кьачIана ваГIаралъи. ЦIакъ ссин бохумо вуго Кьурач Билкьисид абураб жоялъухги ГIенехь`ого къамо рукьги хьумо ячумо гьей машинада жано кIусле гьуна, дилъа мун дурго эбел-энссодаги вассалаздаги ссе щвелулаха йигомамо. Кьурачиасса ГIемер хIинкьиялъ Билкьисид бугIумо бугоха богорукьиса бахчумо басумо нус кодохьумо. Машинаа анаго гьубуле жо кьIахуре Билкьисид нус жидерго чохIоль кьале гьубуна. Билкьис лукьарамалъльа Кьурачид хIвеладаймамо хIинкъумо гьей балнисаялъе ячина.¶

¶

Гъеб кинабго хабар лъарамалъгъа Билкъисил эбел-энссод гъей ругъун сахлълегІанги теого росулъе ячина, Къурачидаги абина къІокІа мун гъай йехІина бакІалълъухе къагомамо. Гъедин гъез лъалуниги Къурач Билкъисихеги эмо къІасса лълъвамамо гъердена вагІана, рукъе йегъемамо абина багІана. КигІан абуниги Билкъис разилъинаро. Билкъис разилъ`амалъгъа Къурачид лъимал нахе рачина, ва Билкъисихе ращанаро. Лъимал жедеассаго ракъІалъи Билкъисе цІакъ захІмалъина. Гъедин ана ункъо-щуго моцІ. Гъокъо-гъокъо лъималазухе шагъаралъгъе ана ягІана Билкъисги. Билкъисе гъакъделего Къурачил эбел-энссод гъессе росулъгъа соги чІчІужо абумо ягІана, амма Къурач разилъинаро. Рукъи чІчІемо хІал къана бугІумо Билкъисги хІалтІоэ лълъвана. Гъелъул цІияб гІумро байбихъина, гІемерал гъудулзаби рагІана. Гъедин ана моцІцІалги рагІана. ЦІакъ ракІ бащаре, берсине Билкъис хІалтІоаги киназего йокъина.¶

¶

Со пуланаб къоялъ гъелъгъа садах хІалтІино вас гъей гъарле Билкъисил энссохе ана. ХІалтІоасса щвараго Билкъисида бисина энссод, нужер бухгалтер ХІажи вегъумо вагІа мун гъарле мамо. Билкъисид абина энссода, дир лъималги темо дун къІокІа чиясе анорола мамо. Гъедиамо бакъІа-бакъІад гІумро гъубуле Билкъисидаги Къурачидаги чІаргІун хъана. Къурач вегъина гъезул рукъе, Билкъисил эбел-энссода ссе гІодо накбиги чІвамо къІаса лъвамамо гъарина гъеза. Билкъисил ракІ бохІина Къурачих, ва эбелги гІебеха гъеого ругал лъималги ракІалъе щвемо гъей гъесса садах шагъаралъе ана. Гъеасса нахен Къурачид гъеб мукилдайги тана. Эбелги ячумо энссо вегъарамалъа лъималги цІакъ рохина. Гъеасан нахе Къурачги Билкъисги гъоркъо щибниги хабарги хъеого рохалиа гІумро гъубуна руго.¶