АВТОМАТИЧЕСКАЯ РАЗМЕТКА НЕПРЕДИКАТИВНЫХ ФОРМ В ЧУКОТСКИХ ТЕКСТАХ

Игнатенко Дарья Игоревна

1. Введение[1]

В настоящее время корпусный подход к исследованию языка всё чаще рассматривается как одна из важнейших методологий для изучения различных языковых явлений (см. например [Leech 1992]). Использование корпусов значительно упрощает поиск материала, необходимого для решения конкретных лингвистических задач, позволяет применять к языковым данным квантитативный анализ, и таким образом корпуса позволяют выстраивать одновременно точные и интересные с теоретической точки зрения модели языка. Однако несмотря на стремительное развитие этого направления в лингвистике, текстовые корпуса существуют для относительно небольшого количества языков: некоторые ареальные и генетические группы языков остаются неохваченными, что представляет проблему для применения корпусного подхода в типологии.

Одна из важнейших составляющих корпуса – парсер, который производит автоматический морфологический анализ словоформ из текста. Программная система UniParser, разработанная Т. А. Архангельским, решила проблему универсализации технической стороны грамматического анализа. Тем не менее для применения данного универсального парсера к конкретному языковому материалу необходимо составление формального описания исследуемого языка.

Целью настоящей работы является описание именного словоизменения и словообразования чукотского языка в формате системы автоматического морфологического анализа UniParser [Архангельский 2012]. Кроме имён также рассмотрены и некоторые другие периферийные части речи. В качестве основы взяты грамматические описания, предложенные в работах [Dunn 1999] и [Скорик 1961, 1977]. Важно отметить, что задачей работы было составить такое описание морфологии чукотского языка, по которому производилась бы корректная разметка текстов, однако не было цели составить полно-

[1] Статья подготовлена в ходе проведения исследования (17-05-0043) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2017-2018 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

стью корректное лингвистически теоретическое описание грамматики языка, что соответствует концепции UniParser [Архангельский 2012: 29].

1.1. Структура слова в чукотском языке

В данном разделе будут рассмотрены некоторые особенности морфологии и фонологии чукотского языка, релевантные для вышеозначенной задачи. Чукотский язык по грамматическому строю относят к агглютинативным и инкорпорирующим языкам. Большинство показателей являются суффиксальными, однако также встречаются префиксы и циркумфиксы.

Для фонетики чукотского характерна гармония гласных по подъёму, все гласные делятся на две серии: сильную (а, э, о) и слабую (э, и, у); шва (ы) может сочетаться с гласными любой серии. Чукотский сингармонизм устроен так, что морфема, содержащая гласный сильной серии, вызывает унификацию по серии во всём слове или инкорпоративном комплексе. Таким образом, и корневые, и аффиксальные морфемы, содержащие слабые гласные, имеют два варианта, выбор между которыми происходит в зависимости от фонетического контекста [Скорик 1961: 37]. Общая структура слога чукотского языка — (C)(C)V(C). Большие стечения согласных или гласных избегаются. Как правило, в сочетания более чем из двух согласных вставляется эпентетическая шва: *вэтгав* 'речь' + *к* (LOC) > *вэтгав-ы-к* 'в речи'. При стечении гласных на стыках морфем предшествующий гласный утрачивается: *га* + *орвыма* > *горвыма* 'с-нартой', *к,ора* + *эн* > *к,орэн* 'олений'. В стыковой ситуации CCV+V оба гласных сохраняются: *умкуум* 'лес', *гамгаорвык* 'на каждой нарте'. В интервокальном положении щелевые факультативно утрачиваются: *аалёмка* < *авалёмка* 'неслышно', *н,ээкык* < *н,эйэкык* 'дочь'. В некоторых случаях на стыках морфем при стечении согласных и внутри морфемы при утрате гласного происходят ассимилятивные и диссимилятивные процессы (см. подробнее [Володин, Скорик 1997:]).

2. Описание лексики и грамматики чукотского языка в формате UniParser

2.1 Лексика

В качестве основы для словаря лексем был использован распознанный словарь [Молл, Инэнликэй 1957]. В файле, содержащем неглагольные лексемы (lexemes.txt), на данный момент всего 3436 лексических единиц, из них 2717 имён существительных, 43 местоимения, 584 наречий, 21 союз и 32 междометия. Каждой лексеме приписана информация о частиречной принадлежности, её перевод на русский язык, перечислены

различные алломорфы основы и словоизменительные парадигмы, свойственные каждой из лексем. Ниже приведён пример описания одной лексемы:

```
-lexeme
 lex: яраңы
 gramm: N
 stem: .яра.//.ра.|.яра.|яраңы.//.ра.|яра.
 trans_ru: 1) яранга; 2) дом
 paradigm: N-obl
 paradigm: N-nom-0
 paradigm: N-pl
```

В формате UniParser информация о лексеме оформляется в виде пар «свойство-значение», каждая пара записывается на отдельной строке [Архангельский 2012: 33]. В поле `lex` записана лемма (соответствует номинативу единственного числа). В `gramm` указана часть речи, что никак не влияет непосредственно на то, с какими показателями сочетается лексема, но может быть использовано, например, при поиске по корпусу. В строке `stem` для каждого из существительных перечислено четыре основы, которые разделены символом «|». Каждая из основ автоматически получает номер, начиная с 0: нулевая — исходная основа с нейтральной гармонией гласных, первая — основа с гласными сильной серии, вторая соответствует основе номинатива единственного числа и последняя — основа, к которой присоединяется показатель номинатива множественного числа; каждая из флексий, которые описаны в следующем разделе, имеет ссылку на конкретную основу и от неё порождает словоформу. Через двойной слэш указаны варианты одной основы (например, в случае если основа существительного меняется при инкорпорации в глагольный комплекс, а также при наличии разных диалектных вариантов одной основы). Точками у основы обозначены места, куда может присоединяться флексия. При каждой именной лексеме стоит ссылка на парадигмы косвенных падежей и именительного падежа единственного и множественного числа; парадигмы отличаются для имён и местоимений.

2.2. Словоизменение

В файле с описанием словоизменения (paradigms.txt) перечислены парадигмы, внутри которых представлен набор взаимоисключающих флексий. Для каждой флексии указана грамматическая информация, которая будет приписана словоформе, содержащей эту флексию. Для примера приведём здесь описание показателя дательного падежа, который входит в парадигму косвенных падежей существительного `N-obl`:

```
-flex: <1>.гты
```

```
 gramm: dat
 gloss: DAT
 regex-stem: [аеёиоуыэюя]\.$
-flex: <1>.эты
 gramm: dat
 gloss: DAT
 regex-stem: [бвгджзйкқлмнңпрстфхцчшщ'ъь]\.$
```

В первой строчке после `-flex:` в угловых скобках записан номер основы, к которой присоединяется показатель, точкой обозначено место основы, а после неё идёт сам аффикс. В поле `gramm` записана грамматическая информация о словоформе, содержащей данный показатель. В строке `gloss` указана глосса аффикса. В данном случае показатель дательного падежа имеет два фонетически обусловленных алломорфа — поэтому в парадигме `N-obl` есть два элемента, имеющих одинаковую глоссу DAT. После `regex-stem:` записано регулярное выражение, задающее условие для основы, к которой может присоединяться флексия.

Итоговый список парадигм представлен в Таблице 1.

Таблица 1. Словоизменительные парадигмы (paradigms.txt)

Название парадигмы		Часть речи
-paradigm: N-nom-0	Номинатив единственного числа с нулевым показателем	Имя существительное
-paradigm: N-nom-n	Номинатив единственного числа, маркированный суффиксом *-н*	
-paradigm: N-nom-ŋə	Номинатив единственного числа, маркированный суффиксом *-ңы*	
-paradigm: N-nom-lgyn	Сингулятив	
-paradigm: N-pl	Номинатив множественного числа	

-paradigm: N-obl	Косвенные падежи для неодушевлённых предметов	
-paradigm: N-НА	Косвенные падежи для имён, обозначающих человека	
-paradigm: N-pers	Категория лица	
-paradigm: PPron-case	Падежи	Личные местоимения
-paradigm: PPron-poss	Притяжательные формы	
-paradigm: adv		Наречия
-paradigm: intj		Междометия
-paradigm: conj		Союзы
-paradigm: part		Частицы

2.2.1. Имя существительное

В чукотском языке в словоизменительной парадигме имён существительных представлены три грамматические категории: падеж, число и лицо. Стоит заметить, что лицо в некоторых работах рассматривается не как грамматическая категория имени, а как форма предикативов [Володин, Скорик 1997], но несмотря на это в рамках поставленной задачи кажется более целесообразным описывать данные показатели как именную категорию.

Существует два набора падежно-числовых показателей для существительных чукотского языка. В зависимости от сочетаемости с ними существительные можно поделить на три группы: имена, обозначающие неодушевлённые предметы (в том числе, вопросительные и указательные существительные), склоняющиеся по первому типу, имена собственные, которые склоняются по второму типу, и последняя группа — имена нарицательные, обозначающие человека, которые могут сочетаться с показателями как первого, так и второго типа с определённым различием в значении [Muravyova, Daniel, Zhdanova 2001]. В рамках данной работы было принято решение исключить из рассмотрения имена собственные, так как представляется затруднительным занесение всех необходимых основ в словарь.

Для первого склонения — имён нарицательных, не обозначающих человека, число формально различимо только в именительном падеже. Показатель множественного числа одинаков для всех имён, единственное же число номинатива маркируется несколькими разными способами, и поэтому были созданы отдельные парадигмы для единственного и множественного числа именительного падежа. В чукотском существует четыре стратегии маркирования на существительном номинатива единственного числа:

- использование исходной основы без каких-либо специальных показателей
- маркирование суффиксом; всего таких суффикса два: архаичный *-ӈы*, который сохранился только для некоторых частотных имён, и суффикс *-н*, который используется со всеми производными именами, а также с некоторыми непроизводными
- частичная или полная редупликация основы
- использование частично супплетивной основы — встречается очень редко

[Dunn 1999: 105]

Для имён, образующих формы именительного падежа единственного числа с помощью суффиксов, заведены две отдельные парадигмы N-nom-n и N-nom-ŋə, все остальные случаи описаны как супплетивные основы с нулевым показателем — парадигма N-nom-0. Таким образом, в файле с лексиконом каждой именной лексеме приписана парадигма N-pl и одна из парадигм N-nom-n, N-nom-ŋə или N-nom-0.

В парадигме N-obl представлено 9 косвенных падежей для имён нарицательных, показатели которых приведены в Таблице 2:

Таблица 2. Парадигма N-obl

Падеж	Показатель
ERG/INST	-та, -тэ, -а, -э
LOC	-к, -кы, -ык
ABL	-йпы, -эпы, -гыпы
DAT	-гты, -эты
ORIENT	-ыгъет, -ыгйит, -гъет, -гйит
DSGN	-но, -ну, -о, у
COM	га-_-ыма, га-_-ма, г-_-ыма, г-_-ма

Таблица 2. Парадигма N-obl

ASSOC	га-_-та, г-_-та, га-_-а, г-_-а, гэ-_-тэ, г-_-тэ, гэ-_-э, г-_-э
PRIV	а-_-ыка, а-_-ка, -ыка, -ка, э-_-ыкэ, э-_-кэ, -ыкэ, -кэ

Большинство показателей имеет алломорфы с гласными обеих серий, такие сингармонические варианты отсутствуют только у аффиксов аблатива, датива и ассоциатива, (эти показатели являются морфемами сильного ряда), и у локатива, который не содержит гласных, кроме швы. Многие показатели также имеют чередования, вызванные различными морфонологическими процессами. Все фонетические чередования, падения гласных и случаи диссимиляции были представлены как алломорфы показателей. Выбор между алломорфами происходит в зависимости от основы, условие для которой задаётся регулярным выражением при каждой флексии.

Как уже было упомянуто ранее, в данной работе рассматриваются нарицательные одушевлённые имена, но не имена собственные. Такие существительные могут сочетаться со всеми показателями перового склонения, но для них также возможны падежно-числовые формы, образованные с показателями для второго склонения. Различие в значении состоит в том, что сочетаясь с показателями второго типа имена становятся более определёнными, а формы множественного числа приобретают значения ассоциативной множественности.

Таблица 3. II склонение

Падеж	SG	PL
NOM	= I склонение	-нти, -нтэ, -ынти, -ынтэ
ERG	-нэ, -на, -ынэ, -ына	-рык, -ырык
LOC	-нэ, -на, -ынэ, -ына	-рык, -ырык
DAT	-на, -ына	-рыкы, -ырыкы
ABL	= I склонение	-ргыпы, -ыргыпы
ORIENT	= I склонение	-рыгйит, -ырыгйит, -рыгъет, -ырыгъет

Образование форм номинатива единственного числа одинаково для обоих склонений, особые формы у одушевлённых имён появляются в косвенных падежах. В отличие от первого склонения, показатели данного типа различают единственное и множественное число во всех падежах (см. Таблицу 3).

Аффиксы, совпадающие с 1 склонением, не были повторно описаны в парадигме N-HA. Показатели эргатива и локатива формально совпадают и различимы только при анализе синтаксиса предложения, поэтому соответствующим аффиксам была присвоена глосса ERG/LOC.

Некоторым именам для реализации значения единственности необходимо наличие показателя сингулятива. Для этого показателя была заведена парадигма N-nom-lgyn. У суффикса сингулятива, имеющего в глубинной форме вид *-*лӈ*, есть три поверхностных реализации: -*лг, -ылг* и -*лың*.

При присоединении показателей лица от существительных образуются формы со значением ‘быть X’. Формы третьего лица единственного и множественного числа совпадают с номинативом, поэтому в парадигме N-pers описываются только показатели первого и второго лица.

Этот набор флексий присоединяется к нулевой основе существительного, и, как и большинство описанных ранее, имеет алломорфы, которые выбираются в зависимости от фонетического контекста — на выбор влияет серия гласных в основе и её окончание. Таким образом, парадигма N-pers содержит 12 элементов: 4 показателя и 3 алломорфа для каждого из них

2.2.2. Личные местоимения

Личные местоимения чукотского языка имеют только одну словоизменительную категорию — падеж. Падежная парадигма для местоимений порождается по правилам, несколько отличным от существительных, поэтому для них был заведён отдельный набор флексий (-paradigm: PPron-case). Для примера приведём ниже склонение местоимения ‘мы’ (см. Таблицу 4).

Таблица 4. Падежная парадигма местоимения *мы*

ABS	muri
ERG	mor-ɣə-nan
LOC	mur-ək
ABL	mor-ək-ajpə

ALL	mor-ək-aɣtə
DAT	mor-ək-ə
ORIENT	mur-ək-eɣjit
COM	ɣe-mur-ək-e
ASSOC	ɣa-mor-əɣ-ma
DSGN	mur-ək-u

[Spencer 1999]

Местоимение имеет два алломорфа основы с гласными разных серий, а также в некоторых формах перед падежным показателем появляется тематический суффикс *-(ы)к/-(ы)г*. Формы с тематическими показателями для удобства описания представлены как алломорфы основ, и, таким образом, в файле lexemes.txt местоимение *мы* записано следующим образом:

```
-lexeme
 lex: мури
 gramm: PRON
 stem: мури.|моргы.//морыг.|.морык.//.мурык.
 trans_ru: мы
```

Нулевая основа соответствует номинативу, первая — основа местоимения для эргатива и ассоциатива, вторая — основа всех остальных косвенных падежей. Остальные местоимения (1sg − *гым*-, 2sg − *гыт*-, 3sg − *ын*-, 2pl − *тур*-, 3pl − *ыр*-) были описаны аналогично.

Парадигма PPron-case была описана в формате, аналогичном флексиям для имён существительных.

2.2.3. Другие части речи

Наречия, междометия, союзы и частицы не имеют словоизменительных категорий, и потому для них для всех парадигмы состоят только из одного элемента — нулевой флексии (`-flex: <0>`.), необходимой для того, чтобы эти лексические единицы были распознаны парсером.

Одним из спорных мест грамматики чукотского языка является категориальный статус прилагательных [Володин 1997]. Так как группа слов, иногда описываемых как прилагательные, в аттрибутивном значении встречается инкорпорированной, а вне ин-

корпоративного комплекса получает глагольное оформление, описание этого материала выходит за рамки настоящей работы.

2.2.4. Последовательности флексий

Многие из перечисленных выше парадигм являются взаимоисключающими, но это верно не для всех комбинаций грамматических значений — так, например, за показателем сингулятива обычно следует суффикс *-н* номинатива единственного числа. Чтобы парсер распознавал подобные цепочки флексий, необходимо эксплицитно определить связь между парадигмами. Для последовательности показателей сингулятив-номинатив описание устроено следующим образом: в парадигму номинатива была добавлена дополнительная флексия, в которой указана ссылка на N-nom-lgyn. Отдельная флексия в парадигме понадобилась потому, что сингулятив — морфема сильного ряда и образует формы только от основы имени с гласными сильного ряда, и таким образом присоединяется к основе, отличной от основы номинатива. Данный элемент парадигмы именительного падежа приведён ниже.

```
-flex: <1>.<.>ын
  gramm: nom,sg
  gloss: NOM
  paradigm: N-nom-lgyn
```

Точкой в угловых скобках (<.>) в первой строке обозначено место присоединения следующей флексии (в данном случае — показателя сингулятива).

2.3. Словообразование

Файл derivations.txt содержит словообразовательные модели. «Описание деривации выглядит в целом так же, как описание лексемы, за тем исключением, что вместо перечисления свойств — основы, грамматических значений и т. п., в деривации перечисляются правила, позволяющие получить значения этих свойств для деривированной лексемы» [Архангельский 2012: 71-76]. Для примера приведём здесь описание одного из суффиксов аугментатива:

```
-deriv-type: aug2
 lex: <0>[.]йң
 stem: [.]йң.
 gloss: AUG
```

В первой строке после `deriv-type`: записано название деривации, в следующей описан вид леммы деривированной лексемы, в строке `stem` — правило, по которому

образуются её основы, а в последней глосса, которая будет приписана словообразовательному аффиксу. Точкой в круглых скобках обозначается место основы.

Алломорфы деривационных аффиксов при их наличии записаны как разные вхождения в файле, но с одинаковой глоссой.

По умолчанию производные лексемы наследуют все словоизменительные парадигмы исходной лексемы. Однако некоторые из деривированных лексем образуют номинатив не так, как лексема-источник — для таких словообразовательных моделей новые парадигмы эксплицитно прописываются.

Чтобы задать, к каким лексемам может быть применима та или иная деривация, ссылка на неё приписывается к словоизменительной парадигме из paradigms.txt, которая в свою очередь связана с набором лексем из lexemes.txt.

Все описанные деривации могут сочетаться с именными основами, поэтому ссылки на них были приписаны к парадигме N-obl, которая сочетается со всеми существительными. Некоторые из словообразовательных показателей являются транскатегориальными и из описанных в рамках данной работы частей речи могут сочетаться также и с наречиями. Ссылки на последние были занесены, помимо N-obl, в парадигму adv.

3. Заключение

3.1. Тестирование морфологического анализатора

Из текстового файла с 18 154 токенами был составлен частотный список, в котором получилось 14 194 словоформы, включая предикативные. Из этих данных парсером было проанализировано и размечено около 38%.

Также для проверки корректности работы анализатора вручную была размечена сказка из сборника [Беликов 1979] (180 токенов, 119 словоформ). По результатам сравнения в разметке парсера ошибок не было найдено. Размечено 55% токенов — все непредикативные формы без инкорпорации.

3.2. Дальнейшие перспективы

Для получения более точных результатов разметки необходимо более детально изучить проблему нерегулярности орфографии и фонетических чередований.

Кроме того, парсер может послужить вспомогательным инструментом для пополнения материала словаря — среди нераспознанных слов оказались такие, которых не нашлось в словаре, но при этом некоторые словарные входы являлись довольно прозрачными деривациями от этих основ.

Также предстоит решить проблему описания инкорпорации, которая не была затронута в рамках настоящей работы.

Библиография

Архангельский 2012 — Т. А. Архангельский. *Принципы построения морфологического парсера для разноструктурных языков: дис. канд. филол. наук*, М., 2012

Беликов 1979 — Л. В. Беликов (ред.). *Лымн'ылтэ: чукотские народные сказки и предания*. Магадан: Кн. Изд-во, 1979.

Володин 1997 — А. П. Володин. Чукотско-камчатские языки // *Языки мира. Палеоазиатские языки*. М.: Индрик, 1997. С. 12–22.

Володин, Скорик 1997 — А. П. Володин, П. Я. Скорик. Чукотскиӣ язык // *Языки мира. Палеоазиатские языки*. М.: Индрик, 1997. С. 23–39.

Молл, Инэнликэй 1957 — Т. А. Молл, П. И. Инэнликэй. *Чукотско-русский словарь*. Л., 1957

Скорик 1961 — П. Я. Скорик. *Грамматика чукотского языка. Ч.1: Фонетика и морфология именных частей речи*. М.; Л., 1961

Скорик 1977 — П. Я. Скорик. *Грамматика чукотского языка. Ч.2: Глагол, наречие, служебные слова*. Л.: АН СССР, 1977.

Dunn 1999 — M. Dunn. *A Grammar of Chukchi*. Ph. D. Diss., Australian National University, 1999

Leech 1992 — G. Leech. Corpora and theories of linguistic performance // *J. Svartvik (ed)., Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, pp. 105–122. Berlin: Mouton de Gruyter, 1992

Muravyova, Daniel, Zhdanova 2001 — I. A. Muravyova, M. A. Daniel, T. JU. Zhdanova. *Chukchi language and folklore in texts collected by V.G.Bogoraz*. M.: 2001. (Unpublished).

Spencer 1999 — A. Spencer. *Chukchi grammar*. http://privatewww.essex.ac.uk/~spena/Chukchee/CHUKCHEE_HOMEPAGE.html