МОРФОЛОГИЧЕСКОЕ ОПИСАНИЕ ЯЗЫКА И ГЛОССИРОВАНИЕ ТЕКСТА: ЗАДАЧИ ЛИНГВИСТА И АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ (НА МАТЕРИАЛЕ СЕЛЬКУПСКОГО ЯЗЫКА)

А. А. Егорушкин
МГУ, ОТиПЛ
egorushkin@mtu-net.ru

А. И. Кузнецова
МГУ, ОТиПЛ
kuzn@kuznec.mccme.ru

Ключевые слова: морфологическая разметка, корпус текстов, языки малочисленных народов, глоссирование.

В настоящей работе даётся описание системы автоматического глоссирования и попытки применения данной системы к текстам на селькупском языке. Объектом глоссирования является максимальная единица морфологического описания (словоформа или аналитическая словоформа), а задачей – выделение в её со-ставе значимых единиц (морфем) и приписывание значения последним. Глосси-рование является практической реализацией морфологического описания языка, и поэтому общепринятым стало, что лингвистическое описание языка должно со-провождаться отглоссированными текстами. Однако часто наблюдается несоот-ветствие морфологического описания и глоссирования. В целях избежания этого была разработана автоматическая система глоссирования селькупского текста, основанная на его морфологическом описании.

1. Введение

В данной статье речь пойдёт о компьютерной системе, предназначенной для работы с корпусами текстов на малых и исчезающих языках. Статья состоит из двух частей. В первой части излагаются основные задачи системы и общие принципы её устройства. Во второй части описывается, каким образом данную систему можно применить к реальной задаче. В качестве полигона был выбран селькупский язык (тазовский диалект).

2. Архитектура системы

2.1. Круг задач

Данная система (будем называть её UniGloss) разрабатывалась в первую очередь для автоматической морфологической разметки корпусов текстов на языках малочисленных народов. Языки малочисленных народов – это потенциальных объект интереса для полевых лингвистов. Таким образом, система UniGloss предназначена главным образом для полевых лингвистов. Одной из основных отличительных черт корпусов текстов на "малых" языках является то, что в качестве минимальной единицы разметки в них

выбирается морфема, а не словоформа[1]. Это обуславливает и сложность автоматической разметки таких корпусов. Так, для разметки корпуса текстов, например, на русском языке, необходим морфологический парсер, который

1) лемматизирует данную словоформу (то есть приводит её к словарной форме),
2) определяет грамматические характеристики данной словоформы.

Но для построения корпуса текстов на "малых" языках необходим более сложный морфологический анализатор, который выполнял бы следующие функции:

1) морфемное членение данной словоформы,
2) приписывание значения выделенным морфемам.

Следовательно, и опираться подобный морфологический анализатор должен на более полное и более совершенное лингвистическое описание морфологии языка, которое должен создать сам лингвист, работающий над корпусом. Однако создать такое лингвистическое описание изолированно от языкового материала не представляется возможным, поэтому все этапы работы над описанием должны верифицироваться на разрабатываемом корпусе. Таким образом, этап разработки формального морфологического описания языка является наиболее важным этапом работы над корпусом и не должен быть недооценен при разработке компьютерных систем подобного типа.

Итак, система UniGloss предназначена для выполнения следующих задач:

- разработка формального морфологического описания исследуемого языка,
- тестирование описания на корпусе текстов,
- полная морфологическая разметка корпуса на основе сделанного описания.

2.2. Компоненты системы

Как явствует из задач, система состоит из следующих основных блоков:

- блок морфологического описания,
- блок морфологического анализа,
- блок разметки корпуса,
- блок работы с корпусом (индексация, составление конкордансов, формирование поискового запроса, поиск, работа с выборкой и т.п.).

Все эти компоненты объединены в единую интегрированную среду, но фактически могут использоваться в других приложениях, поддерживающих технологию COM (независимое[2] использование компонентов становится важным, если есть необходимость публиковать корпус текстов на электронных носителях или в Internet). Взаимосвязи выделенных блоков приведены на рис. 1. Более подробное описание всех компонентов приведено в следующих разделах.

---

[1] Выбор в качестве минимальной единицы разметки словоформы, а не морфемы для таких языков, как английский, французский, немецкий, русский, определяется традицией и сравнительной простотой морфологии данных языков.

[2] Здесь слово "независимое" употреблено в техническом смысле. Безусловно, все эти компоненты являются логически связанными (например, блок морфологического анализа использует данные из блока морфологического описания). Но они не привязаны жёстко к интегрированной среде и поэтому могут быть портированы в ту среду, в которой лингвист привык работать (например, Microsoft Office).

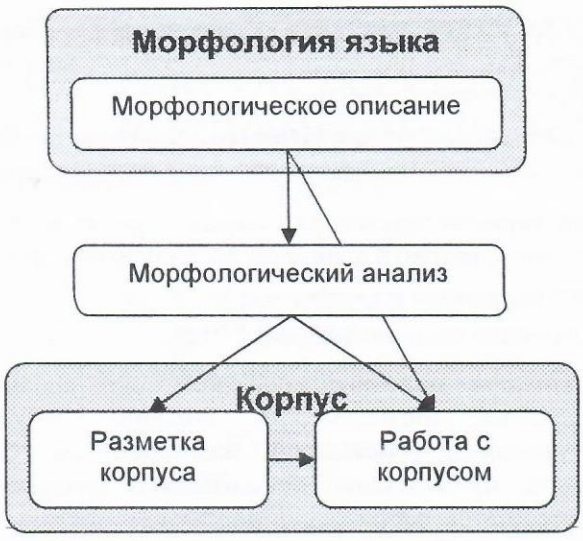

Рис. 1. Взаимосвязи между компонентами системы UniGloss

2.3. Блок морфологического описания

Блок морфологического описания предназначен для создания формального описания морфологии исследуемого языка, которое будет использоваться морфологическим анализатором.

2.3.1. Теоретическая база. Морфологическое описание заключается в перечислении всех морфем, встречающихся в языке, всех алломорфов (морфов), соответствующих морфемам, и правил комбинации морфем в словоформу (синтактика). Вместо перечисления всех алломорфов допустимо задание правил их порождения морфонологическими правилами.

Эта структура соответствует задаче, которая стоит перед морфологическим анализатором. Отметим здесь ещё раз специфику этой задачи. Перед морфологическим анализатором (в нашем случае) не стоит задача определения морфосинтаксических характеристик словоформы. Вместо этого основной его задачей является членение словоформы на морфемы и приписывание им значения[3]. Следовательно, морфологические теории, которые исходят из морфосинтаксических свойств и/или отрицают понятие морфемы[4], здесь оказываются не пригодными.

2.3.2. Компоненты морфологического описания. Таким образом, морфологическое описание состоит из следующих компонентов:

- спецификация объектной модели словаря,
- словарь,
- морфологические правила,
- фонология, фонологические и морфонологические правила.

---

[3] Предполагается, что значения морфосинтаксических характеристик словоформы могут быть выведены из результатов морфемного членения с точностью до синтаксического контекста. В некоторых случаях это может быть достигнуто объединением значений морфем, в других – введением более сложных правил (см. [Кибрик 1997, 27-29]). Пока этот вторичный по отношению к морфемному членению компонент в системе отсутствует. Но для полноценной разметки он необходим и будет добавлен в следующую версию системы.

[4] К ним относятся в первую очередь теории, представленные в [Matthews 1991], [Anderson 1992], [Stump 2001].

Взаимосвязи всех компонентов морфологического описания отражены на рис. 2. Более подробно каждый компонент рассмотрен в следующих разделах.

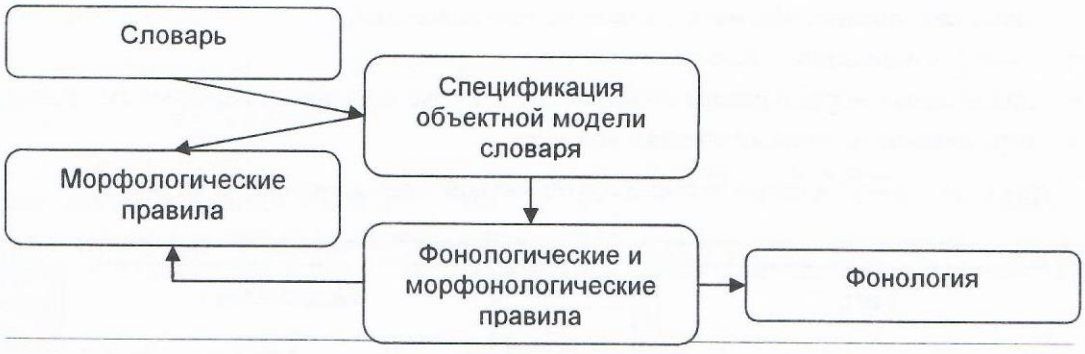

Рис. 2. Взаимосвязи компонентов морфологического описания

2.3.3. Спецификация объектной модели словаря и словарь. Спецификация объектной модели словаря является связующим звеном между всеми компонентами морфологического описания, а также между блоком морфологического описания и блоком работы с корпусом. Спецификация объектной модели задаёт

- типы морфем,
- свойства, соответствующие всем типам,
- описания свойств.

Каждая морфема словаря должна принадлежать к одному типу. Тип определяет структуру свойств морфемы и свойств всех морфов, принадлежащих данной морфеме. Свойство – это поле в описании морфемы или морфа, которое должно быть заполнено лингвистом, редактирующим словарь, в соответствии с описанием этого свойства. Свойство морфа заполняется индивидуально для каждого морфа, принадлежащего данной морфеме.

Все свойства делятся на "системные" и определяемые пользователем. К первым относятся те свойства, без заполнения которых морфологический анализатор не сможет работать, или те свойства, изменение которых переопределяет работу анализатора по умолчанию. Свойства, определяемые пользователем, – это те свойства, которые вводит в описание сам лингвист.

К системным свойствам морфемы относятся:

- глосса (ярлык, соответствующий значению морфемы),
- комментарий (расшифровка ярлыка, указывает на место данной морфемы в системе языка; подсказка для лингвиста, который потенциально будет пользоваться разрабатываемым корпусом).

К системным свойствам морфа относятся:

- означающее морфа (цепочка символов или тип нуля[5]),
- морфологический тип (корневой морф или один из следующих: префикс, суффикс, инфикс[6]),

---

[5] Система UniGloss позволяет работать в описании с нулевыми морфами. Можно выделить несколько типов нулей: "парадигматические" (соответствует знаку с нулевым означающим - Ø); несегментные (когда какое-либо значение выражено несегментно, например, тоном); пустые (когда определённая позиция в словоформе может быть незанята – обычно используется при описании в рамках полиаффиксной морфологии, чтобы сократить число шаблонов); редупликации (сегментное наполнение определяется морфонологическими правилами); конверсии.

- если морф является частью циркумфикса, то позиция его в данном циркумфиксе,
- символ-разделитель морфов (обычно морфы в составе словоформы отделяются друг от друга дефисом, но иногда для некоторых типов морфов используют другой знак; так, клитики обычно отделяют знаком равенства),
- символ-разделитель глосс (например, -, =, +, :, . и др.)[7],
- для нулевых морфов режим отображения в строке значений (подробнее см. далее),
- ограничения на фонологический контекст.

Структура словаря и спецификации его объектной модели отражены на рис. 3.

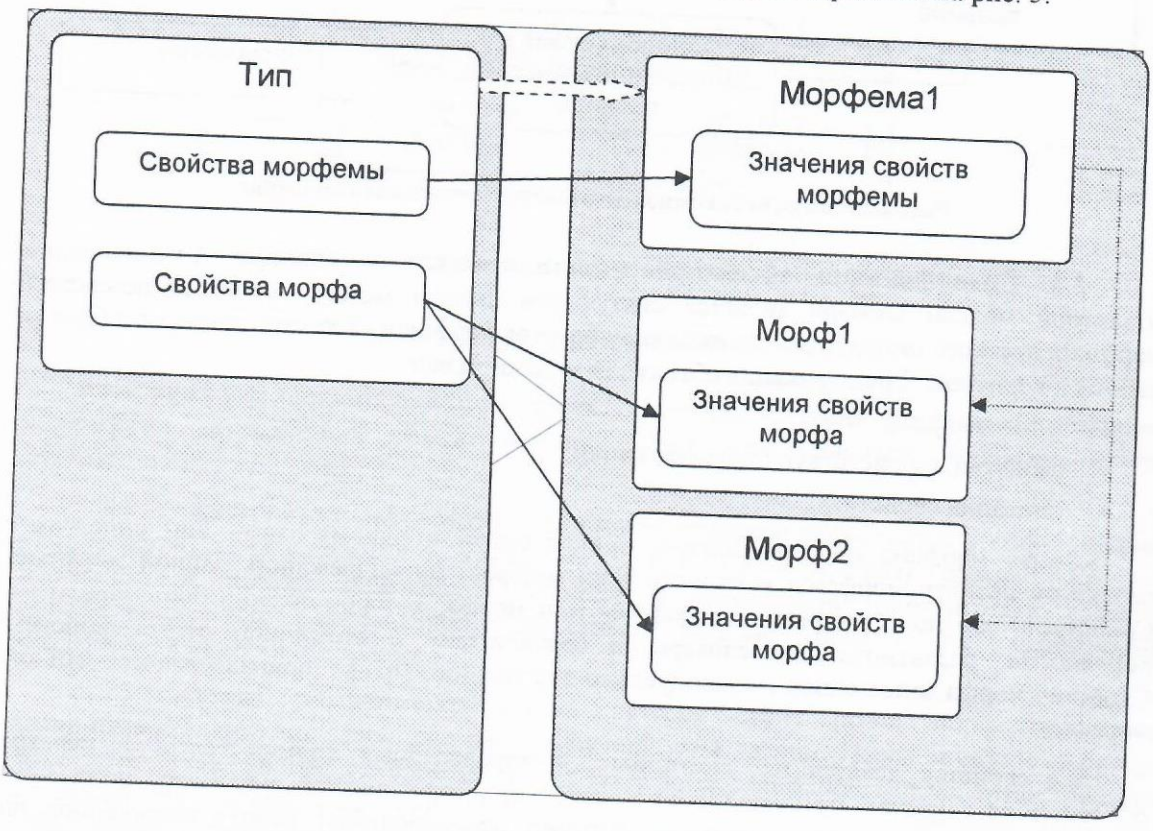

Рис. 3. Структура словаря и объектной модели словаря

Пунктиром изображено отношение принадлежности морфемы типу, а морфа – морфеме. Обычные стрелки показывают, что значения свойств должны соответствовать их описанию.

Введение в морфологическое описание компонента спецификации позволяет, прежде всего, сделать прозрачной его структуру и, кроме того, даёт возможность обращаться к словарю из других блоков системы.

2.3.4. Морфологические правила. Блок морфологических правил задаёт ограничения на сочетаемость морфем в словоформе. Без этого компонента можно обойтись[8], но тогда

1) будет порождаться большое количество неверных результатов анализа,
2) будет невозможно использовать нулевые морфы,
3) морфонологические правила потеряют смысл.

---

[6] Позиция для вставки инфикса задаётся морфонологическими правилами.
[7] Об интерпретации символов-разделителей см. [Lehmann 1982].
[8] Система Shoebox (http://www.sil.org/computing/shoebox/) обходится без блока морфологических правил, что делает её морфологический анализатор практически бесполезным для языков со сложной морфологией.

На данном этапе развития системы морфологическое правило отражает структуру словоформы и ограничения на её части, записанные в виде сравнения значений свойств морфем или морфов. Обобщённо правило выглядит следующим образом[9]:

```
тип1-тип2-тип3:
тип1.свойствоX=тип2.свойствоY;
тип3.свойствоZ='значение'
```

Согласно этому правилу, словоформа считается правильной, если она состоит из трёх морфем, которые принадлежат типу1, типу2 и типу3 соответственно. Причём, значение свойстваX первой морфемы должно пересекаться[10] со значением свойстваY второй морфемы, а значение свойстваZ третьей морфемы должно пересекаться со значением 'значение'.

Каждому морфологическому правилу может быть приписан упорядоченный список морфонологических правил.

2.3.5. Фонологический компонент. В задачи фонологического компонента входит:

1) порождение алломорфов морфемы по правилам (вместо задания их списком),
2) изменение алломорфов в словоформе под влиянием контекста,
3) проверка фонотактических условий правильности словоформы.

Все правила фонологического блока делятся на автоматические, или собственно фонологические (которые применяются независимо от морфологического контекста), и неавтоматические, или морфонологические (которые применяются только в определённых морфологических контекстах). В морфонологических правилах доступны значения всех свойств морфем и морфов.

Скажем несколько слов о формальной модели фонологии, используемой в системе. В генеративной фонологии используются контекстно-зависимые грамматики [Karttunen 1991]. В двух-уровневой фонологии, разработанной финским лингвистом Киммо Коскенниеми, используются конечные автоматы [Antworth 1990; Karttunen et al 1992]. Однако общей чертой этих двух моделей является то, что в качестве единицы алфавита выбирается фонема – последовательность символов, которая далее не разлагается на более мелкие "атомы". Категории фонем, такие как, например, гласные, согласные, сонорные и т.п., моделируются множеством, которому приписано имя. Следствие, вытекающее из данного подхода, отмечается во многих публикациях, посвящённых формальным фонологическим теориям: алфавит сильно разрастается и словарный облик морфа очень сильно отличается от его облика в составе словоформы. В статье [Karttunen et al 1992] приведён яркий пример: лексическое представление финской словоформы *otin* 'я взял' будет выглядеть как *otTalIIn*, где *T*, *aI*, *II* – фонемы с особыми морфонологическими свойствами.

Мы же предлагаем моделировать фонему, считая её не неразложимой последовательностью символов, а объектом некоторого типа (или, что то же самое, –

---

[9] В следующую версию системы планируется добавить возможность использовать в морфологических правилах промежуточные объекты между морфемой и словоформой.

[10] В системе не используется строгое равенство, поскольку все значения свойств являются недоопределёнными, а для них строгое равенство смысла не имеет. Механизм "недоопределённых вычислений" позволяет делать предсказания не только о структуре словоформы, но и о свойствах неизвестных частей этих словоформ. На основе этого механизма в перспективе можно реализовать блок "самообучения" морфологического описания по корпусу. В задачи этого блока будет входить:

1. генерация списка корневых морфов,
2. объединение корневых морфов в морфемы,
3. определение значений свойств корневых морфов.

Таким образом, используя морфологическое описание, по корпусу можно будет получить словарь корневых морфем.

матрицей значений признаков). Категории фонем – тоже объекты, у которых некоторые значения недоопределены. Такой подход позволяет:

1) привести запись означающего морфа к более естественному виду (вместо *a1* и *a2* будет записано *a*, но эти два объекта будут отличаться значениями некоторого признака),
2) единообразно обрабатывать в правилах фонемы и категории фонем,
3) моделировать просодические признаки признаками фонем (например, наличие/отсутствие ударения, границы слогов и т.п. будут признаками фонем).

Фонологические правила должны быть упорядочены. В них допустимо проверять, изменять значения свойств объектов и заменять, удалять и вставлять объекты в определённую позицию.

2.4. Морфологический анализ

Морфологический анализатор членит словоформу на морфемы и приписывает значения выделенным морфемам. Значение берётся из поля Глосса словарной статьи. Если для одной словоформы получается несколько вариантов анализа, то пользователь должен выбрать правильный вариант. Если же не было найдено ни одного варианта или все найденные варианты неправильны, то анализатор генерирует все возможные предсказания. Предсказания генерируются, только если неизвестной является корневая морфема в однокорневой словоформе.

Алгоритм, по которому работает морфологический анализатор, очень похож на алгоритмы, используемые в промышленных морфологических парсерах (см., например, [Сегалович&Маслов 19988]). Используя блок фонологических правил, система генерирует все алломорфы морфем. Далее, по шаблонам словоформ генерируются все "хвосты" – последовательности некорневых морфов словоформы (словоформа за исключением корней). Часть "хвостов" отсекается условиями правил (теми выражениями из множества условий, оба операнда которых принадлежат данному хвосту). Приписанные морфонологические правила корректируют сегментный облик "хвостов". Все "хвосты" и корневые морфы упорядочиваются для ускорения поиска. Это подготовительный этап, который выполняется при "компиляции" морфологического описания[11].

На стадии анализа словоформы происходит поиск "хвостов" от начала и от конца словоформы и их сопоставление (с учётом правил из шаблона). Для всех получившихся "хвостов" из словоформы вычленяется корень (словоформа минус "хвост"). Если корень в словаре существует, то он подставляется в шаблон и проверяется по условиям. Если же корня не существует, то данный результат рассматривается как гипотетический. Для гипотетического корня все свойства полагаются равными максимально недоопределённому значению, и после применения условий их значения доуточняются.

Данный алгоритм даёт высокую скорость анализа, что очень важно для быстрого обновления корпуса в процессе создания, тестирования и "подгонки" морфологического описания[12].

2.4.1. Результат морфологического анализа. Результатом морфологического анализа данной словоформы является последовательность морфов; для каждого морфа указано значение этого морфа, которое берётся из поля Глосса морфемы, включающей данный морф. Далее этот результат можно преобразовать в несколько форматов.

---

[11] Скорость этапа подготовки морфологического описания к анализу зависит от сложности фонологических правил и от длины "хвостов", но практически не зависит от количества корневых морфем. Поэтому система эффективно работает со словарями больших объёмов (25 000 корней).

[12] Поскольку подробное описание алгоритма не является целью статьи, то многочисленные нюансы работы морфологического анализатора опущены для ясности изложения.

Один из наиболее распространённых в лингвистической литературе форматов выглядит следующим образом:

il-ɛnny-nna-nty[13]

жить-Fut-Ltn-2Sg.S

Первая строка называется строкой текста, а вторая – строкой глосс. В строке текста помещены морфы, разделённые символами-разделителями морфов, а в строке глосс – значения морфов, разделённые символами-разделителями глосс. Порождение результата в данном формате называется глоссированием. Данный формат предназначен для вывода пользователю и для цитирования примеров в публикациях.

Для хранения корпуса используется формат XML.

2.4.2. Обработка нулей. Морфологический анализатор имеет несколько опций, касающихся обработки нулей. Следует различать нулевые морфы – Ø и нулевые морфемы – морфемы, у которых все морфы нулевые. По умолчанию, ни нулевые морфы, ни нулевые морфемы не отображаются в строке текста (строка текста выглядит *iča* вместо *iča-Ø-Ø*). По умолчанию же, нулевые морфы, принадлежащие ненулевой морфеме, отображаются в строке глосс, а нулевые морфы не отображаются. Для всех нулей можно включить режим отображения в строке текста, а для любой нулевой морфемы и любого нулевого морфа можно изменить режим отображения в строке глосс. Доступны следующие опции:

- не отображать в строке глосс: Ича[14],
- отображать в строке глосс через символ-разделитель: Ича:Sg:Nom,
- отображать в строке глосс в скобках: Ича(Sg)Nom,
- отображать в конце строки глосс: Ича(Sg:Nom).

2.4.3. Омонимия при анализе. Омонимия при анализе возникает в тех случаях, когда анализатор генерирует два или более результата анализа для одной словоформы. Омонимия бывает двух типов.

Первый тип – наличие у одной словоформы двух различных членений. Например, морфологический анализ словоформы *ilylī* даёт следующие результаты:

(а) il-y-lī̯ (б) ily-l-ī̯

жить-Prs-2Du жить-Opt1-1Du

Опыт работы над корпусами показывает, что процент омонимии данного типа ничтожно мал.

Второй тип омонимии, возникающей при морфологическом анализе, – наличие у одного членения словоформы двух или более значений. При этом большая часть результатов анализа различается значением одной морфемы. Рассмотрим примеры из селькупского языка. Семантику и/или значения морфосинтаксических характеристик всех трёх словоформ, приведённых далее, можно доуточнить только контекстом.

(1) nätä-nyk

а. девушка-Dat 'девушке'

б. девушка-All 'к девушке'

(2) ī-ta-p

а. взять-Fut-1Sg.O 'я возьму'

---

[13] Здесь и далее: все примеры взяты из тазовского диалекта селькупского языка.
[14] Имя главного героя большинства сказок селькупов.

б. взять-Ltn:Prs-1Sg.O 'я, оказывается, взял'

(3) qoŋ-my

а. вождь-Nom:Px1Sg 'мой вождь'

б. вождь-Acc:Px1Sg '(вижу) моего вождя'

Пример (1) демонстрирует случай, когда два значения совмещены в одном показателе (то есть всегда выражаются этим показателем). Поэтому имеет смысл оставить один ярлык, а второй поместить в поле комментария.

Пример (2) похож на (1) тем, что одним показателем выражены два значения. Однако есть позиции, в которых значения настоящего времени латентива и будущего времени индикатива противопоставлены:

(4) paŋ̨č-čɛnta-k

спуститься-Fut-1Sg.S 'я спущусь'

(5) paŋ̨č-ča-k

спуститься-Ltn:Prs-1Sg.S 'я, оказывается, спустился'

Поэтому объединить эти значения под одним ярлыком – Lat:Prs или Fut – нельзя, но чтобы убрать омонимию при анализе, можно объединить их под ярлыком Fut/Lat:Prs. Таким образом, единственным результатом анализа словоформы (2) будет 'взять-Fut/Ltn:Prs-1Sg.O' (результаты для (5) и (6) при этом не изменятся, поскольку в них нет омонимии).

Пример (3), напротив, единственная позиция, в которой Nom и Acc совпадают. Присвоив данной морфеме ярлык Nom/Acc:Px1Sg, получим единственный результат анализа словоформы (3): 'шкура-Nom/Acc:Px1Sg'.

Таким образом, "морфологоцентрический" подход к разметке, согласно которому некоторые синтаксические и семантические различия следует игнорировать, позволяет значительно сократить омонимию при анализе и, следовательно, повысить эффективность анализатора.

3. Иллюстрации на материале селькупского языка

3.1. Текст

Данный текст был отглоссирован системой UniGloss в автоматическом режиме.

iča-l'	čaptä
Ича-Adj	**сказка**

Сказка об Иче.

1)	iča	i	imaqota	ily-mp-ɔ̄qi̯.
	Ича	и	старуха	жить-Pst2-3Du.S

Ича и старуха жили.

2)	iča	qən-pa	tūty-l'	tō-nty.
	Ича	пойти-Pst2:3Sg.S	карась-Adj	озеро-Ill.Sg

Ича поехал на карасье озеро.

3) qət-pa-ty wərqy tūty-p.
поймать-Pst2-3Sg.O большой карась-Acc

Поймал большого карася.

4) tü-mpa moqynä.
прийти-Pst2:3Sg.S домой

Приехал домой.

5) imaqota-nyk kəty-mpa-ty "wərqy tūt-ap
старуха-Dat.Sg сказать-Pst2-3Sg.O большой карась-Nom/Acc:Px1Sg
yky am-ty.
Neg1 кушать:Imp-2Sg.O

Старухе сказал: «Моего большого карася не ешь.

6) jesľi amm-ɛ̄-nta-l, mat qu-ľč-ɛnta-k".
если кушать-Intens-Fut/Ltn:Prs-2Sg.O я умереть-Intens-Fut-1Sg.S

Если съешь, я умру».

7) onty qənn-ɛ̄j-a mač-o.
сам пойти-Intens-Prs:3Sg.S лес-Ill.Sg

Сам поехал в лес.

8) moqynä čap tü-ŋa imaqota iča-t wərqy tūty-p
домой только прийти-Prs:3Sg.S старуха Ича-Gen большой карась-Acc
įnnä amm-ɛ̄-mpa-ty.
вверх кушать-Intens-Pst2-3Sg.O

Домой пришёл – а старуха Ичиного большого карася съела.

9) iča qu-ľč-a.
Ича умереть-Intens-Prs:3Sg.S

Ича умер.

10) imaqota tɔ̄ pū-ŋa, iča-m įllä taq-ny-ty.
старуха на.той.стороне перейти-Prs:3Sg.S Ича-Acc вниз доставить-Prs-3Sg.O

Старуха на ту сторону переехала, Ичу похоронила.

11) imaqota moqynä pū-ŋa.
старуха домой перейти-Prs:3Sg.S

Старуха назад переехала.

12) ūty-t ɔ̄mt-a.
вечер-Gen сидеть-Prs:3Sg.S

Вечером сидит.

13) niľčyk kət-y-ty "tɔ̄-ľ pɛläq-qyt qorqy šɛnty
так сказать-Prs-3Sg.O на.той.стороне-Adj сторона-Loc медведь свежий
lə-p am-qontō-qo panįč-č-y."
кость-Acc съесть-2/3Sg-Inf спускаться-Ltn-Prs:3Sg.O

Так сказала: «На той стороне медведь, чтобы свежие кости съесть, спустился».

14)	iča	ịnnä	omt-ịj-a	"kun	ē-ŋa?"
	Ича	вверх	сесть-Intens-Prs:3Sg.S	где	находиться-Prs:3Sg.S

Ича вверх сел (=вскочил): «Где (он) есть?»

15)	imaqota	tə̄	pū-ŋa,	iča-m	moqynä
	старуха	на.той.стороне	перейти-Prs:3Sg.S	Ича-Acc	домой
	pU-t-y-ty.				
	перейти-Trans-Prs-3Sg.O				

Старуха на ту сторону переехала. Ичу назад перевезла.

Оригинальный текст и перевод взяты из [Очерки 1993, 24; 63-64].

3.2. Корректность морфологического описания

Морфологическое описание может быть разной степени корректности. Корректность описания определяется следующими факторами:

- как анализируются грамматически правильные словоформы,
- как анализируются грамматически неправильные словоформы.

Самое "совершенное" описание – это то описание, которое

- для любой грамматически правильной словоформы генерируют все верные варианты анализа,
- для любой грамматически неправильной словоформы не генерируют ни одного варианта.

Описание подобного типа практически довольно сложно создать. Оно имеет смысл при работе с литературным языком и литературными текстами.

Менее строгое описание

- для грамматически правильной словоформы должно порождать все правильные варианты анализа,
- не обязательно должно запрещать все грамматически неверные словоформы.

Описание данного типа гораздо проще создать. С практической точки зрения они хорошо применимы для корпусов разговорных текстов.

3.3. Образование падежно-числовых форм имён

Проиллюстрируем корректный тип описания на примере образования лично-числовых форм непосессивного склонения имени в селькупском языке. Для демонстрации возможностей морфологических правил, все алломорфы будут заданы в словаре. Таким образом, в данном описании не будут использованы морфонологические правила. Хотя их использование, безусловно, сделало бы описание более простым.

Согласно [Очерки 1980], имя может иметь несколько основ:

1) основа Nom,
2) основа Gen. В общем случае основа Gen не выводима из основы Nom, однако в большинстве она является производной от основы Nom.
3) усечённая основа Nom (образуется от основ Nom на *y* его усечением). Будем обозначать эту основу NomTr.

4) основа Du (образуется от основ Nom на гласный, кроме *y*, удлинением этого гласного или совпадает с основой Nom, если она на согласный). Имя не может иметь одновременно основы NomTr и Du.
5) Ряд основ, получается из вышеперечисленных по стандартным ассимиляционным правилам (которые задаются фонологическими правилами).

Категория числа

Sg имеет нулевой показатель.

Du образуется от основы Du показателями *–qį, -qį̄*, или нулевым показателем. Также Du образуется от усечённой основы Nom показателями *–ōqį, –ōqį̄, –ō(j). -(ō)qį* образует основы Nom и Gen двойственного числа, *-(ō)qį̄* – основу Du двйственного числа. *–ō(j)* и нулевой показатель могут выступать только в конце словоформы.

Pl образуется от основы Gen показателем *-t* (образует основу Nom Pl) или *-ty* (образует основу Gen Pl).

Категория падежа

Nom не имеет специального показателя и образуется от основы Nom.

Gen образуется путём присоединения к основе Gen показателя *-n/-t*.

Dat образуется от основы Gen показателем *-nyŋ/-nyk/-ny* (только для единственного числа). Или показателем *-kinį* от формы Gen.

Ill образуется от основы Nom на согласный показателем *-ty*, от основы Nom на гласный, кроме y, – показателем *–nty*, от усечённой основы Nom – показателями *–onty* или *o*. Ill образуется только в единственном числе.

Loc образуется путём присоединения к основе Du (или, факультативно, к основе Nom) показателя *-qyn/-qyt*. Также Loc образуется от усечённой основы Nom показателями *-ōqyn/-ōqyt* и *–ōn*. Loc и Ill не образуются от одушевлённых имён.

Остальные падежи образуются по одной из вышеперечисленных моделей.

Таким образом, чтобы корректно (о корректности морфологического описания см. выше) описать данный фрагмент парадигмы имени, необходимы три типа, которые описаны далее.

```
Тип Падеж
    - этому типу принадлежат все падежные показатели.
    Свойства морфемы:
        • Число (образование некоторых форм ограничено только
          единственным числом). Допустимые значения: Sg, Du, Pl.
        • Одушевлённость (локативные падежи не образуются от
          одушевлённых имён). Допустимые значения: +, -.
    Свойства морфа:
        • Основа (от которой образуется данный падеж). Допустимые
          значения: Nom, Gen, Du, NomTr, CaseGen (если от формы Gen)
Тип Число
    - этому типу принадлежат все числовые показатели.
    Свойства морфемы: нет дополнительных свойств.
    Свойства морфа:
```

- Основа1 (к которой присоединяется данный морф). Допустимые значения: Nom, Gen, Du, NomTr.
- Основа2 (результирующая основа, которая получается после присоединения показателя). Допустимые значения: Nom, Gen, Du, NomTr.

Тип Имя

- этому типу принадлежат все именные основы.

Свойства морфемы:

- Одушевлённость. Допустимые значения: +, -.

Свойства морфа:

- Основа (какую основу представляет данный морф). Допустимые значения: Nom, Gen, Du, NomTr.

Все морфы Падежа и Числа являются суффиксами. Все морфы Имени являются корневыми. Ниже приведён фрагмент словаря[15]:

Морфема:
Тип: Число
Глосса: Sg
Отображение в глоссах: не отображать
Морф:
Вид: Ø (парадигматический)
Основа1: Nom
Основа2: Nom
Морф:
Вид: Ø (парадигматический)
Основа1: Gen
Основа2: Gen
Морф:
Вид: Ø (парадигматический)
Основа1: Du
Основа2: Du
Морф:
Вид: Ø (парадигматический)
Основа1: NomTr
Основа2: NomTr

Комментарий: Sg, в отличие от Du и Pl, не переопределяет значение свойства Основа именной основы. Поэтому равенство свойств Основа1 и Основа2 имеет смысл "Показатель Sg, присоединяясь к основе, образует основу того же типа". В следующих версиях системы "прозрачность" морфемы

[15] Словарь и описания типов приведены в текстовом виде. Однако система предоставляет графический интерфейс для редактирования словаря и типов.

по отношению к некоторому свойству можно будет указывать непосредственно в морфологических правилах.

Морфема:

Тип:	Число
Глосса:	Du
Морф:	
Вид:	qi̯
Основа1:	Du
Основа2:	Nom, Gen
Морф:	
Вид:	qī̯
Основа1:	Du
Основа2:	Du
Морф:	
Вид:	j
Основа1:	Du
Основа2:	Nom
Левый контекст:	ā
Правый контекст:	#[16]
Морф:	
Вид:	Ø (парадигматический)
Основа1:	Du
Основа2:	Nom
Левый контекст:	ā
Правый контекст:	#
Разделитель глосс:	:
Отображение в глоссах:	через символ-разделитель
Морф:	
Вид:	ōqi̯
Основа1:	NomTr
Основа2:	Nom, Gen
Морф:	
Вид:	ōqī̯
Основа1:	NomTr
Основа2:	Du
Морф:	
Вид:	ōj
Основа1:	NomTr
Основа2:	Nom

[16] Знак конца словоформы.

```
        Правый контекст:          #
    Морф:
        Вид:                      õ
        Основа1:                  NomTr
        Основа2:                  Nom
        Правый контекст:          #
Морфема:
    Тип:                          Число
    Глосса:                       Pl
    Морф:
        Вид:                      t
        Основа1:                  Gen
        Основа2:                  Nom
    Морф:
        Вид:                      ty
        Основа1:                  Gen
        Основа2:                  Gen
Морфема:
    Тип:                          Падеж
    Глосса:                       Nom
    Число:                        любое значение
    Одушевлённость:               любое значение
    Морф:
        Вид:                      Ø (парадигматический)
        Основа:                   Nom
        Отображение в глоссах:    не отображать
Морфема:
    Тип:                          Падеж
    Глосса:                       Gen
    Число:                        любое значение
    Одушевлённость:               любое значение
    Морф:
        Вид:                      n
        Основа:                   Gen
        Правый контекст:          гласный, сонорный согласный, #
    Морф:
        Вид:                      t
        Основа:                   Gen
        Правый контекст:          шумный, неносовой сонорный, #
Морфема:
```

```
    Тип:                    Падеж
    Глосса:                 Dat.Sg
    Число:                  Sg
    Одушевлённость:         любое значение
    Морф:
        Вид:                ny
        Основа:             Gen
    Морф:
        Вид:                nyk
        Основа:             Gen
    Морф:
        Вид:                nyŋ
        Основа:             Gen
Морфема:
    Тип:                    Падеж
    Глосса:                 Dat
    Число:                  любое значение
    Одушевлённость:         любое значение
    Морф:
        Вид:                kiŋi
    Основа:                 CaseGen
Морфема:
    Тип:                    Падеж
    Глосса:                 Ill.Sg
    Число:                  Sg
    Одушевлённость:         -
    Морф:
        Вид:                nty
        Основа:             Nom
        Левый контекст:     гласный, не y
    Морф:
        Вид:                onty
        Основа:             NomTr
    Морф:
        Вид:                ty
        Основа:             Nom
        Левый контекст:     согласный
    Морф:
        Вид:                o
        Основа:             NomTr
```

```
Морфема:
    Тип:                    Падеж
    Глосса:                 Loc
    Число:                  любое значение
    Одушевлённость:         -
    Морф:
        Вид:                qyn
        Основа:             Du, Nom
    Морф:
        Вид:                qyt
        Основа:             Du, Nom
    Морф:
        Вид:                ōqyn
        Основа:             NomTr
    Морф:
        Вид:                ōqyt
        Основа:             NomTr
    Морф:
        Вид:                ōn
        Основа:             NomTr
Морфема:
    Тип:                    Имя
    Одушевлённость:         +
    Морф:
        Вид:                iča
        Основа:             Nom, Gen
    Морф:
        Вид:                ičā
        Основа:             Du
Морфема:
    Тип:                    Имя
    Одушевлённость:         -
    Морф:
        Вид:                mačy
        Основа:             Nom, Gen
    Морф:
        Вид:                mač
        Основа:             NomTr
```

Морфологические правила:

```
Имя-Число-Падеж:
    Имя.Основа=Число.Основа1;
    Число.Основа2=Падеж.Основа;
    Падеж.Число=Число.Глосса;
    Падеж.Одушевлённость=Имя.Одушевлённость

Имя-Число-Падеж_1[17]-Падеж_2:
    Имя.Основа=Число.Основа1;
    Число.Основа2=Падеж_1.Основа;
    Падеж_1.Глосса='Gen';
    Падеж_1.Основа='Nom';
    Падеж_2.Основа='CaseGen';
    Падеж_2.Число=Число.Глосса;
    Падеж_2.Одушевлённость=Имя.Одушевлённость
```

Данное морфонологическое описание является корректным по отношению к "нормированному" варианту селькупского языка. В нём не учтены некоторые формы, появляющиеся в разговорных текстах.

Примеры морфологического анализа:

iča	iča-t	iča-nyk	ičatynyk	iča-t-kinį	ičaty	ičanty[18]
Ича	Ича-Gen	Ича-Dat.Sg	???	Ича-Gen-Dat	???	???
	Ича-Pl					

mačy	mač-ōqį	mač-ōqį̄-qyt	mačyty	mačy-qyt	mač-o	mačy-n	mač-ōn
лес	лес-Du	лес-Du-Loc	???	лес-Loc	лес-Ill.Sg	лес-Gen	лес-Loc

4. Состояние проекта

Проект UniGloss находится в стадии разработки. Созданы макетные варианты блоков системы и объединены интегрированной средой. В 2002 году проект был поддержан Лицеем Информационных Технологий (www.lit.msu.ru), после чего началась разработка полноценной и законченной версии системы.

5. Список сокращений в глоссах

Acc	аккузатив (винительный падеж)
Adj	адъективная репрезентация существительных
All	аллатив (падеж)
Dat	датив (дательный падеж)
Du	двойственное число
Fut	будущее время
Gen	генитив (родительный падеж)
Ill	иллатив (падеж)

[17] Символ _ отделяет имя типа от имени объекта данного типа в правиле.
[18] Как уже отмечалось выше, посессивные формы не учитываются.

Imp	императив (повелительное наклонение)
Inf	инфинитив
Intens	интенсивно-перфектная совершаемость
Loc	локатив (падеж)
Ltn	латентив (наклонение)
Neg1	отрицание, используемое в императиве
Nom	номинатив (именительный падеж)
O	объектное спряжение
Pl	множественное число
Prs	настоящее время
Pst2	повествовательное прошедшее время
Px	посессивное склонение
S	субъектное спряжение
Sg	единственное число
Trans	показатель переходности

Литература

1. Anderson S.R. A-Morphous Morphology. Cambridge University Press. 1992.

2. Antworth E.L. PC-KIMMO A two-level Processor for Morphological Analysis. Summer Institute of Linguistics. Occasional Publications in Academic Computing. 1990.

3. Karttunen L. Finite-State Constraints // Proceedings of the International Conference on Current Issues in Computational Linguistics. June 10-14. 1991. Universiti Sains Malaysia. Penang. Malaysia (http://www.xrce.xerox.com/competencies/content-analysis/fsCompiler/articles/fsc-91/fsc91.html19).

4. Karttunen L., Kaplan R.M., Zaenen A. Two-Level Morphology with Composition. // Proceedings of Coling 92. International Conference on Computational Linguistics. Vol. I 141-148. July 25-28. 1992. Nantes. France (http://www.xrce.xerox.com/competencies/content-analysis/fsCompiler/articles/coling-92/coling92.html).

5. Lehmann C. Directions for interlinear morphemic translations. // Folia Linguistica XVI. 1982. pp 199-224.

6. Matthews P.H. Morphology (second edition). Cambridge University Press. 1991.

7. Stump G.T. Inflectional Morphology. A Theory of Paradigm Structure. Cambridge University Press. 2001.

8. Кибрик А.Е. Иерархии, роли, нули, маркированность и "аномальная" упаковка грамматической семантики. // Вопросы языкознания. №4. 1997. сс. 27-57.

9. Очерки по селькупскому языку (тазовский диалект). Том 1. М., 1980.

10. Очерки по селькупскому языку (тазовский диалект). Том 2. М., 1993.

11. Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для неописанных в словаре слов. // Труды международного семинара Диалог'98 по компьютерной лингвистике и её приложениям. Том 2. 1998. сс. 547-552.

Morphological Description and Glossing: Problem for Linguist and Task for Computer System (based on the materials from the Selqup language)

A. A. Egoroushkin, A. I. Kouznetsova

Keywords: morphological tagging, text corpus, glossing.

The report describes the basic principles of interacting between the morphological description of language and the automatic text glossing system. These principles are illustrated by applying the system tuned up by morphological description to the texts of the north dialect of the Selqup language.

---

[19] WWW-ссылки проверялись 31 марта 2002 года.