Г.А. Дырхеева

О ГРАММАТИЧЕСКОЙ ПРОБЛЕМАТИКЕ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ БУРЯТСКОГО ТЕКСТА

Статистическая обработка данных в области языка и литературы характеризуется множеством разнообразных подходов и решений. Наиболее очевидным объектом лексического и стилистического изучения является словарь. Частотный словарь какого-либо произведения или автора, подвергающий обработке текст с первой и до последней страницы и, таким образом, выявляющий все многообразие слов, использованных автором, представляет широкие перспективы для всестороннего изучения словаря данного автора, его стиля. "Изучение же методами (приемами) статистики индивидуальных речевых стилей открывает большие возможности в исследовании речевых стилей и речевого новаторства, речевых потоков и направлений, связанных с деятельностью виднейших представителей национальной словесной культуры"[1].

Автоматическая обработка бурятского текста проводится впервые. Исходным материалом для этого послужили прозаические произведения Х. Намсараева[2]. Результаты обработки будут представлены в виде "Частотного словаря прозаических произведений Х. Намсараева". Естественно, что новый для данного языка подход, тем более что он относится к формальному типу обработки, вызвал ряд вопросов в традиционной грамматике и подчеркнул те ее малоразработанные стороны, с которыми нам пришлось столкнуться. На предварительном, домашинном, этапе работы грамматическая проблематика была затронута дважды: при индексации текста и при формулировании правил формирования словника.

Разработанная нами система индексации имеет своей целью получение информации о морфологических классах слов, а также устранение лексико-грамматической омонимии. Необходимо отметить, что индекс* приписывался каждому словоупотреблению. Лексиче-

* Данная система в своей основе близка существующим системам индексации индоевропейских текстов, а также системе, разработанной для тюркских текстов[3].

ские омонимы в основном не были разграничены, например: хурэхэ 'доходить' и 'остывать', та 'Вы' и 'вы' и др., поскольку машине такие операции не под силу. Однако в некоторых случаях, например: аша 'внук' и 'милость, услуга', турэ 'свадьба' и парное к слову нэрэ 'правление' и др., они были подразделены уже после машинной обработки вручную.

Вопрос об отнесении того или иного слова к какой-либо части речи, к сожалению, непростой даже для такого теоретически разработанного языка, как русский. В бурятском языкознании, лингвистические традиции которого не так богаты, этот вопрос стоит еще более остро, поэтому, естественно, что в своей работе мы, в основном, опирались на Академическую грамматику бурятского языка[4], а также на ряд работ, в которых затрагиваются данные проблемы[5]. В процессе индексации текста мы старались учитывать всю совокупность признаков данного слова, однако слабая дифференциация многих из них, наличие переходных случаев*, различные подходы к решению некоторых грамматических проблем часто затрудняли решение вопроса, поэтому в таких случаях приходилось либо обращаться к компетентной теоретической литературе, либо решать такие случаи коллегиально.

Выделенные грамматические классы обозначены следующим образом:

С – имя существительное: а) собственно имена существительные, б) субстантивированные прилагательные, обозначающие "признак, не соотносимый с определенным конкретным предметом, и функционирующие в значении абстрактных существительных типа русских "величина", "темнота", "злоба" и пр."[6], например: тарган 'полнота', ута 'длина', ябаган 'пешеход' и т.д., в) субстантивированные причастия с суффиксами -гша, -аша: абарагша 'защитник', угэгшэ 'податель', худэлмэрилэгшэ 'работник', буудаашан 'стрелок', наадаашан 'игрок' и т.д., г) субстантивное употребление имен прилагательных, т.е. их падежных форм: мууе, мууешье, муухайтаа, муухайда и т.д., а также случаи, когда за-

---

* "Известно, что одни части речи могут переходить в другие, не изменяя своих внешних показателей... Подобных переходных частей речи в монгольских языках очень много... Иногда признаки перехода одних частей речи в разряд других частей речи бывают не так явственны. В таких случаях единственным критерием может явиться значение слова, которое изменяется при переходе слова из одной категории в другую"[7].

висимое существительное опущено: бодо мал → бодо 'крупный рогатый скот'.

П – имя прилагательное; этим индексом мы пометили также существительное в функции определения: мяхан 'мясной', абын 'отцовский', зурхэнэй 'сердечный', мунгэн 'серебряный' и др.; что касается имен с суффиксами -тай и -гуй, находящихся "как бы на промежуточной полосе между именем прилагательным и существительным"[8], то здесь мы следовали принципу, выдвинутому Ц.Б. Цыдендамбаевым: "адъективированными образованиями" являются имена с отвлеченным и переносным значениями[9].

М – местоимение; Ч – имя числительное; Н – наречие; состав данного класса слов еще четко не установлен, многие формы находятся на стадии перехода в эту группу; в нашем случае к наречиям отнесены также прилагательные, существительные, деепричастия в функции обстоятельств и усилительные слова типа яар, эгээн, хамха, угаа, тон и др.; формы, образованные от различных частей речи с помощью суффикса орудного падежа -гаар, -аар: аймшагтайгаар 'страшно, ужасно', номгохоноор 'смиренько, тихонечко', эдэбхитэйгээр 'активно', уже фактически закрепились в значении наречий.

Г – глагол; Д – деепричастие; Р – причастие.

Классификация глагольной группы слов в основном опирается на морфологический признак, в ее основу положены принципы составления картотеки "Толкового словаря бурятского языка", разработанные в отделе языкознания Бурятского филиала СО АН СССР. Согласно этим принципам, падежно оформленные причастия имеют индекс Р. Попытка выделить различные грамматические значения глагола гэ- 'говорить' оказалась безрезультатной, поскольку нет четко обоснованной дифференциации функций и значений данного глагола.

Т – частица; некоторую информацию для предполагаемого в дальнейшем синтаксического анализа текста могут дать введенные дополнительные обозначения: ТМ – модальные, ТС – уступительные, ТУ – утвердительные, ТВ – вопросительные, ТО – отрицательные частицы; к частицам нами отнесены также связки типа бэлэй, близкие по своим функциям к частицам.

Л – послелог; сюда мы отнесли также случаи типа хадаа, болбол, гээшэ и т.д., являющиеся показателями подлежащего, поскольку введение дополнительных индексов на связки, показатели под-

лежащего значительно увеличило бы количество индексов.

З – союз; Ж – междометие; О – модальное слово; несмотря на то, что вопрос о существовании модальных слов в бурятском языке фактически не разработан, мы попытались выделить их в качестве самостоятельного класса слов, опираясь при этом на работы[10].

К – имя собственное; географические названия, клички людей и животных, названия статей, книг, должностей, титулов и т.д., пишущиеся с большой буквы, выписываются со строчной: Верховно → верховно-П, Герой → герой-С, Главнокомандалагша → главнокомандалагша-С.

И – изобразительное слово; выделено Ц.Б. Цыдендамбаевым[11]; предложение Л.Д. Шагдарова[12] рассматривать изобразительные слова в сочетании с глаголом гэ- мы считаем не совсем верным. Конечно, чаще всего они встречаются в союзе с этим глаголом, но возможны и другие варианты, а, кроме того, при автоматической обработке текста невозможно, например, представить случай типа "гули... гули... гул... гул... гэжэ" (звукоподр. птичьему щебету), поскольку приписывание каждому слову гули глагола гэжэ изменит частотную картину. Поэтому изобразительные слова даны нами как самостоятельные словоупотребления.

В – вводное слово; данная группа слов немногочисленна, хотя в функции вводного слова могут выступать многие части речи, мы выделили лишь те вводные слова, которые было трудно квалифицировать иначе: нэгэдэхеэр 'во-первых', гурба-гэбэл 'в-третьих' и др.

В качестве единицы лингвистического анализа выбрано словоупотребление. Под словоупотреблением понимается любая последовательность букв и знаков между двумя пробелами. При таком подходе автоматически разрываются многие устойчивые словосочетания. Чтобы отчасти восполнить этот пробел искусственно, знаками = и ж были соединены некоторые устойчивые сочетания слов:

а) парные слова очень употребительны в художественной литературе, они образуются из любой части речи, а также составляются из лексических единиц разных диалектов или языков (бурятского и русского). В бурятской лингвистической литературе существуют разногласия по поводу способа образования парных слов, их классификации[13]. Наш принцип выделения парных слов

близок к точке зрения Г.Д. Санжеева[14]. Он считает, что парные слова - результат сложения отдельных знаменательных, близких по значению слов (смежных, синонимичных или контрастных), с целью создания нового понятия, для выражения категории множественности, для лучшего понимания слов разных языков или диалектов. Несомненно, при данном словообразовании полученное сочетание оказывается эмоционально, экспрессивно более нагруженным, чем каждый компонент парного слова в отдельности. Примеры парных слов: шашаха шалиха 'болтать' (оба компонента имеют одинаковое значение, т.е. два синонимичных по значению слова объединены с целью выражения категории множественности), ажал худэлмэри 'работа' (оба компонента имеют одинаковое значение), убһэ хулһан 'корма' (букв. 'сено' и 'камыш', т.е. смежные по значению слова создают новое понятие), тухэл шарай 'облик' (букв. 'вид' и 'выражение'), эд товар 'товар' (слова разного происхождения сложены для лучшего понимания слов разных языков) и др. Компоненты парных слов, представляющие собой единое смысловое целое, соединены двумя черточками: айл-аймаг 'община, люди', ядаха-дутаха 'терпеть нужду', амиды-мэндэ 'живой и здоровый' и т.д.

б) из разряда парных слов нами исключаются следующие виды двусловных образований, выделенные различными авторами[15]:

I) сочетания, в которых первый или второй компонент не имеет самостоятельного употребления: дагжа һураха 'обессилеть всем телом', хара торюун 'тьма-тьмущая';

2) сочетания, в которых оба компонента, взятые в отдельности, не имеют самостоятельного употребления: абар-табар 'редкий', убдэл субдэл 'остатки';

3) сочетания, образованные из числительных или составные числительные: хоёр зуу 'двести', арбан табан 'пятнадцать';

4) сочетания, образованные из наречий места, времени, образа действия, меры и т.д.: дутуу ядуу 'слегка', хайра гамгуй 'без жалости', даб гэһээр 'сразу', нюла нюла 'вяло, неохотно', хуха нуга 'пере-';

5) сочетания, образованные из междометий, служебных и звукоподражательных слов: хай даа 'едва ли', һайн даа 'хорошо', халагни халаг 'вот горе-то какое';

6) сочетания с усилительными частицами типа саб, шад, хаб,

образующие качественные прилагательные: саб сагаан 'белый-пребелый', шад улаан 'ярко-красный'.

Все эти случаи объединены в группу устойчивых парных* сочетаний, соединенных при перфорации и выдаче на печать знаком ж. В случае, когда сочетания һэб-һаб (говорится о дуновении ветерка), даар-дархи 'в беспорядке нагроможденный', ялаг-ёлог 'мерцать, поблескивать' встречаются без черточки, мы их также соединяем знаком ж, т.е. һэб-һаб и һэбжһаб считаются разными словоформами. Поскольку парные слова и слова со знаком ж фиксируются как одно слово, их частота при подсчете общего количества словоупотреблений удваивается.

в) знаком % при перфорации выделялись различные фразеологизмы, например: зурхөө ухэхэ 'отказываться от мысли' (букв. умирать от сердца), хухэ буха 'синица' (букв. синий бык), ухэр нюдэн 'черная смородина' (букв. бычий глаз), газар дуулаг, гахай шагнаг 'об этом история умалчивает' (букв. пусть слушает земля, пусть обнюхивает свинья), урагшаа һанаатай 'прогрессивный' (букв. стремящийся вперед) и др. Эти словосочетания в отличие от слов, соединенных знаками = и ж, не подвергаются статистической обработке как одно целое, а как словоупотребления приводятся отдельным списком в виде словаря фразеологизмов.

В выделении данных фразеологизмов, идиом какой-нибудь определенной классификации не придерживались. Если рассматривать выделенные фразеологические элементы относительно классификации фразеологизмов Ц.Б. Будаева[16], то можно сказать, что в нашем случае выделены в основном фразеологические сращения и единства, а фразеологические сочетания и выражения - отчасти. Бесспорным было выделение пословиц и поговорок.

"Частотный словарь прозаических произведений Хоца Намсараева" в целях более удобного и рационального представления материала несколько видоизменен. Учитывая опыт предыдущих составителей частотных словарей и конкордансов, нами был выбран гнездовой способ представления материала для алфавитно-частотной части словаря, т.е. под каждым заглавным словом приводятся все формы, имевшие место в данном объеме текстов. Каждое

* Исключение составляют междометия типа ээжэ хаажа бай даа сааша 'как бы чего не вышло', включающие три и более компонентов.

словарное вхождение сопровождается информацией об общем количестве встречаемостей словоформ и слов (абсолютной частоте) и количестве встречаемостей в отдельных текстах: I) рассказах, 2) повести "Цыремпил", 3) повести "Нэгэтэ ӈуни", 4) повести "Илалтын туяа", 5) повести "Алтан зэбэ", 6) повести "Эдиршууд", 7) повести "Эжэл гурбан нухэд", 8) повести "Тэршээхэн унаган", 9) романе "Уурэй толон", например:

		I	2	3	4	5	6	7	8	9
адуун–С	39									
адуу–С	3	2								I
адуу–малыемни–С	I	I								
адуу–моридой–С	2		I							I
адуугаа–С	7	4						3		
адууи–С	I5	3		2				I	9	
адуунай–С	4	2								2
адуунайнгаа–С	2									2
адуундаа–С	2							2		
адууншни–С	I	I								
адууӈһаа–С	2		I							I

Приводя такие данные о частоте словоформ по произведениям, мы пытались отразить такую важную характеристику, как распространенность словарной единицы, равномерность ее встречаемости в языке данного писателя.

Преимущества данного гнездового способа представления частотного словаря языка писателя очевидны: помимо общих количественных сведений, лингвист или литературовед, просматривая такой словарь, сможет легко увидеть, какие формы лексемы являются наиболее частотными, для какого произведения они характерны, какие слова являются новыми в данном произведении по отношению к другим трудам писателя.

Сведение словоформ в слова производилось вручную, так как решение такой проблемы машинным способом представляет собой отдельную, далеко не простую задачу даже для традиционной лингвистики.

Заглавные слова располагаются в алфавитном порядке. В случае их омонимии первой по порядку идет глагольная основа, не имеющая индекса, например: хара- 'смотреть', хара-П 'черный'. Парные слова и слова, соединенные знаком ж, входят в гнездо по алфавиту первого компонента сочетания и представлены в на-

чале словарной статьи. Их предлагается рассматривать следующим образом:

ажал-С 'труд'
ажал-байдал-С
ажал-байдалаа-С
ажал-С
ажалай-С и т.д.

абяагуй-П 'беззвучный'
абяа-шэмээгуй-П
абяагуй-П и т.д.

Исходной (словарной) формой для имени существительного считается основа имени, традиционно именуемая формой именительного падежа единственного числа. Выделенные при индексации существительные в функции определения и обстоятельства и обозначенные соответственно как П и Н входят в гнездо имени существительного:

худөө-С 'худон, сельская местность'
худөө-Н 'по-сельски'
худөө-П 'сельский'
худөө-С
худөөдэ-С и т.д.

морин-С 'конь, лошадь'
морео-С
морин-П 'конный'
морин-С
мориной-П
мориной-С и т.д.

В случае, если процесс адъективизации или адвербиализации существительного зашел далеко и слова имеют глубокие семантические расхождения, они считаются разными лексемами, например: хунды-П 'пустой' и хунды-С 'пустота', ногоон-П 'зеленый' и ногоон-С 'трава', соорхой-П 'продырявленный' и соорхой-С 'поляна' и др. Это относится также и к особой группе слов в бурятском языке, обозначающих масти лошадей: буурал-П 'седой, чалый' и буурал-С, хула-П 'саврасый' и хула-С 'савраска'. Формы на -шуул, -дуул, -тан, -хи, -хин, образующие имена существительные с собирательным значением, образуют самостоятельные лексические гнезда.

Имена прилагательные даются в своей исходной форме. Падежные формы имени прилагательного в случае их неполной или, как пишет Т.А. Бертагаев[17], "относительной"* субстантивации включены в гнездо прилагательного:

муу-П 'плохой'

---

* "...отмеченную субстантивацию нельзя признать полной и достигшей качественного завершения, так как она полностью не утверждает прилагательных в роли существительных. Особенно это ясно на фоне тех случаев, когда мы имеем завершившийся процесс субстантивации имен прилагательных"[18].

муу-Н 'плохо'
муу-П
муу-С 'плохое'

Как видно, гнездо прилагательного включает также случаи "относительной" адвербиализации или использования прилагательного в функции обстоятельства, имеющего индекс Н. Так же, как и в случае адъективизации существительных, степень завершенности процесса субстантивизации в каждом конкретном случае разбиралась отдельно.

Суффиксы -хан, -бтар, -шаг, -рхуу, -лиг, -мээр и др., согласно работе[19], считаются словообразовательными, следовательно прилагательные, наречия и числительные с данными суффиксами образуют самостоятельное гнездо так же, как и прилагательные с отвлеченными абстрактными значениями, образованные с помощью суффиксов -тай и -гуй, хотя и здесь иногда трудно провести границу между абстрактными и предметными значениями.

Лексико-грамматические омонимы (омографы) типа адли-Л 'подобен', адли-П 'подобный', адли-Н 'одновременно'; газаа-Л 'вне, за', газаа-Н 'вне, снаружи'; үглөө-С 'утро', үглөө-П 'завтрашний', үглөө-Н 'утром'; юумэ-С 'вещь', юумэ-М 'что-то, нечто', юумэ-ТУ имеют каждый отдельное самостоятельное вхождение.

Глагол, причастие, деепричастие сведены к исходной глагольной корневой основе, не имеющей индекса. Внутри гнезда они также располагаются в алфавитном порядке. Впервые в нашем словаре залоговые формы представляют каждый отдельное самостоятельное гнездо слов, поскольку, как считает Ц.Б. Цыдендамбаев[20]: "Залог в бурятском языке представляет сложную категорию, носящую смешанный, ярко выраженный лексико-грамматический характер"[21].

Производные междометия, образованные от других частей речи, входят в гнездо того слова, производным от которого они являются, например: барьан-Ж от барьан-С 'бедняга, милый', ьайнж даа 'спасибо' от ьайн-П 'хороший, добрый', хөөрхэйждаа-Ж 'эх, бедняга' от хөөрхы-С 'бедняга' и др.

Как уже отмечалось выше, сочетания слов, соединенные знаками = и ж, входят в гнездо по первому компоненту. В случае сложного прилагательного с усилительными частицами: сабжсагаан 'белый-пребелый', шаджулаан-П 'ярко-красный', заглавным словом является частица: саб-Т, шад-Т. Индекс Т используется

также для обозначения заглавного слова непосредственно частиц. Поскольку "в некоторых случаях одни и те же частицы употребляются для выражения разных оттенков"*, в одно гнездо могут входить и разные частицы:

ааб-Т
ааб-ТВ
ааб-ТМ
ааб-ТУ
аабиб-ТМ
аабибди-ТМ
аабта-ТВ и т.д.

Имена собственные имеют самостоятельное словарное вхождение: Эрдэм-К и эрдэм-С (Эрдэм и 'наука, образование').

Разряды числительных даются отдельными словарными статьями. Числительное нэгэн 'один', употребленное в значении неопределенного местоимения 'какой-то' и помеченное индексом М, входит в гнездо числительного.

Супплетивные формы личных местоимений даются по алфавиту их исходных форм в именительном падеже.

Слова, имеющие одинаковое значение и сходное, но не совсем тождественное написание, например: тусхэгэр-тэсхэгэр 'толстый, пузатый', хирхаг-хярхаг 'кайма, кромка', саахар-сахар 'сахар', старшина-таршанаа 'старшина', нөөхи-үнөөхи 'этот самый' и др., считаются разными словоформами, но относятся к одной лексической группе. Эти графические своеобразия объясняются либо диалектными различиями, либо неустоявшейся орфографией, а иногда просто типографскими ошибками.

Произведения Х. Намсараева отличаются выразительностью, богатством словаря, многообразием употребления языковых средств, и поэтому, естественно, что составляя частотный словарь, мы пытались отразить в нем те языковые особенности, которые характерны для Х. Намсараева: своеобразное написание слов (амраг вместо амараг 'любимый'), слова, не зафиксированные в словарях (ороошон и ерээшэн 'входящие'), многочисленное использование муу-II 'плохой' в качестве имени существительного, окказиональное использование алдарта-П 'прославленный' в качестве имени существительного и т.д.

* "...частицы не подвергались специальному анализу ни в одном из монгольских языков"[22].

I Головин Б.Н. Из курса лекций по лингвистической статистике. – Горький, 1966. – С. 21.

2 См.: Намсараев Х. Собрание сочинений в 5 томах. – Улан-Удэ, 1958.

3 См.: Джубанов А.Х. Статистическое исследование казахского текста с применением ЭВМ (на материале романа М.Ауэзова "Абай жолы"): Автореф. на соиск. уч.степ.канд.филол.наук. – Алма-Ата, 1973.

4 См.: Грамматика бурятского языка: Фонетика и морфология. – М., 1962.

5 См.: Алексеев Д.А. Именные части речи в монгольских языках// Вопросы языкознания. – 1955. – № 2. – С. 35–47; Бертагаев Т.А. О морфологическом строе бурятского языка. – М., 1961; Бертагаев Т.А. Проблема классификации частей речи: По материалам монгольских языков// Зап. Бур.НИИК. – Т. 21. – Улан-Удэ, 1956. – С. 36; Дарбеева А.А. О субстантивном употреблении имен прилагательных в бурятском языке// Филология и история монгольских народов. – М., 1958; Санжеев Г.Д. К проблеме частей речи в алтайских языках// Вопросы языкознания. – 1952. – № 6; и др.

6 Дарбеева А.А. О субстантивном употреблении имен прилагательных в бурятском языке// Филология и история монгольских народов. – М., 1958.

7 Бертагаев Т.А. О морфологическом строе бурятского языка. – М., 1961. – С. 30.

8 Цыдендамбаев Ц.Б. Некоторые вопросы составления бурятско-русского словаря// К изучению бурятского языка. – Улан-Удэ, 1969. – С. 122.

9 Там же. с. 123.

10 См.: Балагунова С.С. О языковых средствах выражения модальности в бурятском языке// О содержании и объеме языковой модальности. – Новосибирск, 1982.

11 Цыдендамбаев Ц.Б. Изобразительные слова в бурятском языке// Филология и история монгольских народов. – М., 1958.

12 Шагдаров Л.Д. Изобразительные слова в современном бурятском языке. – Улан-Удэ, 1962.

13 См.: Абашеев Д.А. Об образовании парных слов в бурятском языке: К изучению бурятского языка// Тр. БИОН. – Вып. 6. – Улан-Удэ, 1969. – С. 87–90; Дарбеева А.А. К вопросу о парных словах в бурятском языке// Вопросы литературного бурятского языка. – Улан-Удэ, 1963; Дондуков У.-Ж.Ш., Данчинова Н.Г. К вопросу о парных словах в монгольских языках// О зарубежных монголоведных исследованиях по языку. – Улан-Удэ, 1968; и др.

14 См.: Санжеев Г.Д. Грамматические приемы в монгольских языках// Тр. Ин-та востоковедения. – № 2. – М., 1940.

15 См.: Абашеев Д.А. Об образовании парных слов в бурятском языке: К изучению бурятского языка// Тр. БИОН. – Вып. 6. – Улан-Удэ, 1969; Дондуков У.-Ж.Ш., Данчинова Н.Г. К вопросу о парных словах в монгольских языках// О зарубежных монголоведных исследованиях по языку. – Улан-Удэ, 1968; и др.

16 См.: Будаев Ц.Б. Фразсология бурятского языка. - Улан-Удэ, 1970.

17 Бертагаев Т.А. Проблема классификации частей речи: По материалам монгольских языков// Зап. Бур.НИИК. - Т. 21. - Улан-Удэ, 1956.

18 Там же, с. 63.

19 См.: Грамматика бурятского языка: Фонетика и морфология. - М., 1962.

20 Цыдендамбаев Ц.Б. Грамматические категории бурятского языка в историко-сравнительном освещении. - М., 1979.

21 Там же, с. 85.

22 Шевернина З.В. Функционально-семантическое значение модальных частиц в монгольском языке// Вопросы грамматической системы монгольских языков. - Элиста, 1980. - С. 20.