И.С. Николаев

ПРОБЛЕМЫ МОРФОЛОГИЧЕСКОГО АННОТИРОВАНИЯ КОРПУСА ИЖОРСКИХ НАРОДНЫХ ПЕСЕН

Исследовательская база данных по морфологии ижорских народных песен, которая создана автором предлагаемого доклада, предназначена для получения статистических сведений о грамматических формах, представленных в текстах эпических песен. Материалом для такой базы данных в первую очередь служит корпус текстов «Древние песни финского народа»[1], созданный Финским литературным обществом (Suomalainen Kirjallisuuden Seura) и находящийся в открытом доступе на его веб-сайте[2].

Однако тексты этого корпуса морфологически не аннотированы, поэтому при наполнении базы данных возникает необходимость проводить грамматическую разметку материала. Для бесписьменного и малочисленного ижорского языка не существует практики (полу)автоматического морфологического аннотирования. Тем не менее, для такой работы можно воспользоваться опытом финских исследователей, занимающихся вопросами грамматической разметки тестов на финском языке, который, как и ижорский, является представителем северной подгруппы прибалтийско-финской ветви финно-угорских языков.

Морфологический анализатор финского языка FINTWOL, разработанный финской компанией Lingsoft, Inc.[3], широко распространен в финской корпусной лингвистике и является хорошей основой для построения аналогичного программного модуля для аннотирования текстов ижорских народных песен. Но, очевидно, что возникает ряд сложностей, связанных, прежде всего, со структурой ижорского языка, с некоторыми лингвистическими

---

[1] Suomen Kansan Vanhat Runot

[2] **www.finlit.fi**

[3] **www. lingsoft.fi**

чертами ижорских народных песен, с вариативностью в ижорских диалектах и говорах, а также, не в последнюю очередь, с особенностью записи этих песен финскими исследователями.

В данной статье мы затронем лишь первый из перечисленных аспектов и предложим некий «промежуточный» вариант. Как уже говорилось, ижорский язык входит в ту же подгруппу прибалтийско-финских языков, что и финский, поэтому многие словоформы этих родственных языков совпадают и могут быть проанализированы при помощи FINTWOL. В некоторых текстах песен число таких словоформ достигает 40-50% от их общего количества, что уже значительно повышает скорость предварительного автоматического аннотирования текстов.

Например[1]:

"<miksei>" ‘почему не’ (1)

"miksei" INTERR ADV NEGV V SG3

"<kasva>" ‘расти’ (2)

"kasvaa" V PRES ACT NEG

"kasvaa" V IMPV ACT SG2

"kasvaa" V IMPV ACT NEG SG"

<sängyssä>" ‘в кровати’ (3)

"sänky" N INE S

"<pellervoin>" ‘Пеллерво’ (4)

"pellervo" PROP N INS PL

"<eikä>" ‘и не’ (5)

"ei" COORD C NEGV V SG3 kA

Естественно, после этого этапа эксперт вручную выбирает нужный вариант, например в (2), и проверяет правильность разметки, например в (4).

Словоформы, которые сразу не распознаются и не аннотируются, в основном можно отнести к следующим группам:

---

[1] Примеры из разметки эпической песни «Почему не растет у нас овес» (Miksei kasva meijen kagrat. №1139. Kati-akka, Soikkola, Väärnoja. Volmari Porkka. 1881-83. SKVR III 1).

- отсутствующие в финском языке,
- отличающиеся одной или несколькими фонемами (символами в транскрипции),
- имеющие другие грамматические формы,
- имена собственные.

Такие словоформы морфологическим анализатором FINTWOL не распознаются.

Например:

"<meijen>" ‘наш’ (6)

"<kagrat>" ‘овес’ (7)

"<mättähälläkää>" ‘кочка’ (8)

"<makkais>" ‘лежал бы’ (9)

"<sämpsän>" ‘Сямпся’ (10)

Однако ижорские словоформы второй, например (7) и (8), и третьей группы, например (6), (9) интересны тем, что их соответствие финским носит регулярный характер. Это значит, что возможен алгоритм приведения этих словоформ к «финскому стандарту» и, в конечном итоге, их аннотирование с помощью FINTWOL.

Например:

"<meidän>" (11)

"me" PERS PRON GEN PL

"<kaurat>" (12)

"kaura" N NOM PL

"<mättäälläkään>" (13)

"mätäs" N ADE SG kAAn

"<makaisi>" (14)

"maata" V COND ACT SG3

"maata" V COND ACT NEG

В (6) заменен один суффикс генитива *-jen* на другой *-dän*, в результате получаем правильную аннотацию (11). В (7) заменен согласный *g* на гласный *u*, в результате чего имеем правильную

разметку (12). В (8) опущен согласный *h* в середине словоформы и добавлен согласный *n* на конце, в результате чего получаем (13). В (9) опущен согласный *k* в середине словоформы и добавлен гласный *i* на конце, в итоге имеем два варианта (14).

Разумеется, мы имеем дело с решением *ad hoc*, но создание специального морфологического анализатора для корпуса ижорских народных песен представляется нецелесообразным, так как тексты в этом корпусе имеют значительное диалектное разнообразие и в любом случае требуют трудоемкой ручной работы эксперта. С другой стороны, возможность доработки анализатора FINTWOL для анализа ижорских текстов кажется маловероятной, прежде всего, по причине коммерческого характера самого продукта.

Работа над алгоритмом «нормализации» ижорских словоформ еще не закончена, но он уже сейчас используется в ограниченном виде для аннотирования текстов ижорских эпических песен при их вводе в базу данных.