УДК 811.512.157’366

Linguistic Annotation of Grammatical Categories of Sakha: Nouns

Gavril G. Torotoev* and Sandaara G. Torotoeva
M.K. Ammosov North-Eastern Federal University
58 Belinsky Str., Yakutsk, 677000, Russia

Received 07.11.2018, received in revised form 21.11.2018, accepted 05.12.2018

This paper shows the work to create instruments for linguistic annotation of grammatical categories of Sakha language (Sakha language and Yakut language are full synonyms). It describes the basic inflectional characteristics of Nouns of Sakha language (numbers, personal endings, possessive endings, cases), which are based on Leipzig Glossing Rules. As a result of scientific research (2014-2018) the system of tags was developed, which reflects all word forming potential of the Nouns in the Sakha language, including 247 morphological indicators in its arsenal. It should be noted that the standardized system of morphological tagging of Turkic languages, developed by the Turkologists, is far from perfect, there are various treatments concerning reflection and interpretation of grammatical categories in different Turkic languages. Despite this, the article summarizes constructive and progressive ideas of our colleagues on this matter.

Keywords: linguistic annotation, grammatical categories, Sakha, nouns, numbers, possessiveness, simple declension, possessive declension, diminutive, tags.

Research area: philology.

Citation: Torotoev, G.G., Torotoeva, S.G. (2018). Linguistic annotation of grammatical categories of Sakha: nouns. J. Sib. Fed. Univ. Humanit. soc. sci. DOI: 10.17516/1997-1370-0360.

Introduction

Linguistic annotation of grammatical categories of languages is an up-to-date issue in modern computational linguistics. Artificial intelligence opens an opportunity to get innovative results in theoretical linguistics (acquiring new knowledge about language structure), as well as in applied linguistics (modernization of linguistic research methods, implementation of new technologies for automated language processing).

* Corresponding author E-mail address: torgav@mail.ru

Today, due to the intensive development of computer technologies, there is a need in tagging system for the automatic analysis of electronic corpora of Turkic texts. To improve the effectiveness of comparative studies and acquisition of objective language data as a representative linguistic instrument, it is necessary to apply a standardized morphological tagging system to the corpora of texts in Turkic languages.

A working version of standardized morphological tagging of Turkic languages was accepted in 2014 during UniTurk workshop ("Unification of Grammatical Annotation Systems in the Electronic Corpora of Turkic Languages") in Kazan. The database is built on the morphemic structure of Turkic word forms and is made to reflect structural-semantic model of Turkic languages as precisely as possible. The uniform standard of linguistic information representation opens a unique opportunity for Turkic languages to join the common information space. (Zheltov, 2015: 329).

Problem statement

As a result, it is necessary to develop a tag system which would adequately reflect all grammatical categories of the Yakut language. The work in this area has been in progress for five years to conclude that there are some grammatical categories of the Yakut language that have not been fully reflected in the previous publications. The computer processing of text, which requires complete formalization of knowledge about language and its grammar, reveals some interesting language facts and implicit (hidden) linguistic details, not covered by the classic works by the Yakut scholars.

The computational linguistics researchers have been paying special attention to the inflectional and derivational morphology. Consequently, first it is required to describe and mark all regular inflectional and active derivational indicators of the Yakut language. Secondly, it is necessary to develop rules for the allomorph selection and sandhi rules for automatic word form analysis (morphonological processes in morphemic boundaries; phonetic processes within one word form).

Methods

The study is descriptive. To find the maximum number of inflectional allomorphs of nouns in the Sakha language, the quantitative method was used. As a result of the empirical analysis, nine tables were compiled, forming the basis for the interpretation and reflection of the grammatical categories of nouns in the Sakha language. The

research results may be used for filling lacunas in the existing studies of the Sakha Language.

Discussion

For morphological annotation of grammatical categories of the Sakha language the system of tags based on the Leipzig glossing rules is used. Tags indicating parts of speech in the Sakha are presented in Table 1.

Table 1

Tags	Full term
N	Noun
POSS	Possessive
PRO	Pronoun
NUM	Numeral
ADJ	Adjective
V	Verb
PCP	Participle
CONV	Converb
ADV	Adverb
MOD	Modal word
INTJ	Interjection
CONJ	Conjunction
PART	Particle
POST	Postposition
IMIT	Imitative word

From the point of view of the Sakha language glossing, in this article the grammatical category of nouns was considered. Such inflectional characteristics of the noun as number, case, possessiveness and personality have been carefully analyzed.

1. Number

In the Yakut language, the plural affix *–lar* is represented by 16 forms (Korkina, 1982: 125-126). In the selection of the optimal allomorph, the key role is played by the vowel harmony rules of the Yakut language. Phonetic compatibility of morphemes also depends on assimilation rules (progressive, regressive, progressive-regressive assimilation of consonants) and accommodation. Thus, sandhi rules are developed in accordance with vowel harmony rules, rules of assimilation and accommodation, and demonstrate the sound changes at the morphemic boundaries.

Table 2

Tags	Description	Allomorphs	Morphemes
SG	singular	-	-
PL	plural	-lar/-lor/-ler/-lör -nar/-nor/-ner/-nör -dar/-dor/-der/-dör -tar/-tor/-ter/-tör	-LAr

2. Possessiveness

In the Sakha language, the initial form of the possessiveness category is represented by 58 morphological indicators. These forms are frequently used, as they express various logical relations and connections between objects, that are often different from the concept of possession (Korkina, 1982: 129).

Table 3

Tags	Description	Allomorphs	Morphemes
POSS_1SG	Possessive, 1st person, singular (‘my’)	-m	-m
POSS_2SG	Possessive, 2nd person, singular (‘your’)	-ŋ	-ŋ
POSS_3SG	Possessive , 3rd person, singular (‘his/her/its’)	*after consonants*: -a/-o/-e/-ö *after vowels:* -ta/-to/-te/-tö	-A -тA
POSS_1PL	Possessive, 1st person, plural (‘our’)	-byt/-bit/-but/-büt -pyt/-pit/-put/-püt -myt/-mit/-mut/-müt	-ВУт
POSS_2PL	Possessive, 2nd person, plural (‘your’)	-ɤyt/-ɤit/-ɤut/-ɤüt -xyt-xit/-xut/-xüt -kyt/-kit/-kut/-küt -gyt/-git/-gut/-güt -ŋyt-ŋit/-ŋut/-ŋüt	-ɤУт
POSS_3PL	Possessive, 3rd person, plural (‘their’)	-lara/-lora/-lere/-lörö -nara/-noro/-nere/-nörö -dara/-doro/-dere/-dörö -tara/-toro/-tere/-törö	-LArA

3. Cases in the Yakut language

In the interpretation of grammatical categories and their indication with corresponding tags we relied upon the work of the academician O.N. Boethlingk “About the language of the Yakuts” published in 1851. He registered ten cases in the Yakut language: Casus Indefinitus, Accusativus Indefinites, Dativ, Accusativus Definitus, Ablativ, Lokativ,

Instrumental, Casus Adverbialis, Comitativ, Casus Comparativus (Boethlingk, 1990: 278-285). As it can be seen from the case names, there is no significant difference between the modern terms and those used by O.N. Boethlingk. In the modern Yakut language there are eight cases, Lokativ and Casus Adverbialis are not included into the case paradigm.

Simple declension

There are two types of declension in the Yakut language: simple and possessive (Korkina, 1982: 129-147). In simple declension, all morphemes have 4 allomorphs each, for example: -TA (-ta/-to/-te/-tö).

Table 4

Tags	Description	Allomorphs	Morphemes
NOM	Nominative	-	-
PAR	Partitive	-ta/-to/-te/-tö -la/-lo/-le/-lö -na/-no/-ne/-nö -da/-do/-de/-dö	-TA
DAT	Dative	-ɤa/-ɤo/-ɤe/-ɤö -xa/-xo/-xe/-xö -ga/-go/-ge/-gö -ŋa/-ŋo/-ŋe/-ŋö -ka/-ko/-ke/-kö	-ɤA
ACC	Accusative	*after consonants*: -y/-i/-u/-ü *after vowels:* -ny/-ni/-nu/-nü	-У -nУ
ABL	Ablative	*after consonants*: -tan/-ton/-ten/-tön *after vowels:* -ttan/-tton/-tten/-ttön	-tAn -ttAn
INS	Instrumental	-nan/-non/-nen/-nön	-nAn
COM	Comitative	-lyyn/-liin/-luun/-lüün -nyyn/-niin/-nuun/-nüün -tyyn/-tiin/-tuun/-tüün -dyyn/-diin/-duun/-düün	-LYYn
COMP	Comparative	-taaɤap/-tooɤop/-teeɤer/-tööɤör -naaɤap/-nooɤop/-neeɤer/-nööɤör -daaɤap/-dooɤop/-deeɤer/-dööɤör -laaɤap/-looɤop/-leeɤer/-lööɤör	-TAAɤAr

Possessive declension

In total, simple (88) and possessive declensions (87) have 175 morphological indicators in the Yakut language. It all shows the huge functional capacity of nouns in the Yakut language as a special lexical and grammatical word class.

Table 5

Tags	Description	Allomorphs	Morphemes
NOM	Nominative	See Table 3	*See Table 3*
PAR	Partitive	-yna/-ine/-una/-üne	- УnA*
DAT	Dative	*after consonants*: -ar/-or/-er/-ör *after vowels:* -gar/-ger	-Ap -gAr
ACC	Accusative	-yn/-in/-un/-ün	-Уn
ABL	Ablative	-ttan/-tten	-ttAn
INS	Instrumental	-nan/-nen	-nAn
COM	Comitative	-nyyn/-nuun/-neen *non-literary version:* -naan/-niin	-nУУn
COMP	Comparative	-naaɤap/-neeɤer	-nAAɤAp

*currently out of use.

4. Personal endings of nouns

Nouns in the Yakut language can act as a predicate in sentences. In such cases, predicativity affixes are added to the word root, except the third person singular.

Table 6

Tags	Description	Allomorphs	Morphemes
P_1SG	1st person, singular (‘I am’)	-byn/-bin/-bun/-bün -myn/-min/-mun/-mün -pyn/-pin/-pun/-pün	-BУn
P_2SG	2nd person, singular (‘you are’)	-ҕyn/-ҕin/-ҕun/-ҕün -xyn/-xin/-xun/-xün -kyn/-kin/-kun/-kün -gyn/-gin/-gun/-gün -ŋyn/-ŋin/-ŋun/-ŋün	-ҕУn
P_3SG	3rd person, singular (‘he/she is’)	-	-
P_1PL	1st person, plural (‘we are’)	*after* –LAr: -byt/-bit/-but/-büt	-bУt
P_2PL	2nd person, plural (‘you are’)	*after* –LAr: -gyt/-git/-gut/-güt	-gУt
P_3PL	3rd person, plural (‘they are’)	-lar/-lor/-ler/-lör -nar/-nor/-ner/-nör -dar/-dor/-der/-dör -tar/-tor/-ter/-tör	-LAr

5. Diminutive

Diminutiveness category is an understudied aspect in the Sakha language. Table 7 shows common diminutive affixes -čУk, -čAAn, -kAAn with their allomorphs.

Table 7

Tags	Description	Allomorphs	Morphemes
DIM	Diminutive	-čyk/-čik/-čuk/-čük -čaan/-čoon/-čeen/-čöön -kaan/-koon/-keen/-köön	-čYk -čAAn -kAAn

In addition to these affixes, the Yakut lexical units can consist of fossil affixes such as -yja, -čče, -ka, considered to be of little efficiency at the moment. In Table 8, they are represented downward from the diminutive point of view.

Table 8

Size L	Size M	Size S	Size XS	Size XXS
Lexeme	-yja	-čče	-ka	-čaan
küöl 'lake'	kölüje	kölüčče	kölüke	kölükečeen
ürex 'small river'	ürüje	ürüčče	-	ürüječeen
xolbo 'box'	xolbuja	-	xolbuka	xolbujačaan

6. Derivation

Word-forming potential of nouns in the Sakha language requires a specific approach and a deep study. Without going into details, it should be noted that dozens of productive and non-productive affixes such as -hyt (-syt, -čyt, -djyt, -njyt), -byl (-bil, -bul, -bül), -laŋ (-leŋ, -loŋ, -löŋ), -lta (-lte, -lto, -ltö) and others take part in noun formation in the Sakha language. As an example of derivational affixes, let us consider three frequently used morphemes used to derive verbal nouns.

Table 9

Tags	Description	Allomorphs	Morphemes
AN	Agens noun	-aaččy/-eečči/-oočču/-öočču	-AAččU
VN	Verbal noun	-yy/-ii/-uu/-üü -aahyn/-eehin/-oohun/-ööhün	-UU -AAhYn

Examples of linguistic annotation of nouns

To validate the tag system developed for the linguistic annotation of the word forming potential of nouns in the Yakut language, let us analyze few examples.

(1) xarandaac+(y)nan → xarandaahynan (uruhujduur)

pencil-INS
‘(he draws) with a pencil’
(2) oɤo+ŋ+un → oɤoɤun (köröör)
child-POSS_2SG -ACC
‘(look after) your child’
(3) ubaj+lar+byt+(y)gar → ubajdarbytygar (bierbippit)
brother-PL-POSS_1PL-DAT
‘(we gave) our brothers’
(4) at+lar+ɤyt+(y)naaɤar → attargytynaaɤar (türgen)
horse-PL-POSS_2PL-COMP
‘(faster) than your horses’
(5) ije+te+neen → ijetineen (kelle)
mother-POSS_3SG-COM
‘(he came) with his mother’

Conclusion

During the research (2014-2018), all grammatical categories of nouns in the Sakha language have been analyzed. Through this process, the system, consisting of the conventional symbols (tags) used to reflect the inflectional potential of nouns in the Sakha language, including 247 affixes, has been fully completed.

To enable a computer to automatically analyze texts of any complexity presented in the electronic corpora of the Sakha language, it is necessary to provide standardized tags to all grammatical categories of the Sakha language. The solution of this problem would make it possible to develop new computer programs, such as online translators, automatic text analyzers, speech synthesizers and others.

References

Baker, M.C., Vinokurova, N. (2010). Two modalities of case assignment: Case in Sakha. In *Natural Language & Linguistic theory,* 28, 593–642. DOI: <10.1007/s11049-010-9105-1>.

Boethlingk, O.N. (1990). *O iazyke iakutov* [*About the language of the Yakuts*]. Novosibirsk: Nauka, 646 p.

Kang, D., Torotoev, G. (2016). Morphophonemic derivation of voice in the Sakha language. In *Language, Communication, and Culture. The Journal of the Linguistic Society of the North East,* 3, 66-90.

Korkina, E.I., Ubryatova, E.I., Kharitonov, L.N., Petrov, N.E. (1982). *Grammatika sovremennogo iakutskogo literaturnogo iazyka. Fonetika i morfologiia* [*Grammar of the modern Yakut literary language. Phonetics and morphology*], Moskva, Nauka, 496 p.

Kornfilt, J., Preminger O. (2015). Nominative as no case at all: an argument from raising-to-accusative in Sakha. In *Proceedings of the 9th Workshop on Altaic Formal Linguistics (WAFL 9)*, *MIT Working Papers in Linguistics 76, ed. Andrew Joseph & Esra Predolac,* Cambridge, 109–120.

Levin, T., Preminger, O. (2015). Case in Sakha: are two modalities really necessary? In *Natural Language & Linguistic Theory,* 33, 231–250. DOI: <10.1007/s11049-014-9250-z>.

Torotoev, G.G. (2011). *Funktsional'no-stilisticheskaia differentsiatsiia opredeleniy v sovremennom iakutskom iazyke* [*Functional and stylistic differentiation of the attributive constructions in the modern Yakut language*]. Yakutsk, North East Federal University Publishing and Polygraphic Complex, 148 p.

Torotoev, G.G. (2014). Metod modelirovaniia v issledovanii stikhoobrazuyushchego karkasa olonkho [Method of modeling in the study of Olonkho architectonics]. In *Trudy Kazanskoy shkoly po komp'iuternoy i kognitivnoy lingvistike TEL-2014* [*Proceedings of the Kazan School of computational and cognitive linguistics TEL-2014*]. Kazan', Fen PublishingHouse of the Academy of Sciences, 243-247.

Torotoev, G.G., Nogovitsyna, A.N. (2017). Lingvisticheskoe annotirovanie nakloneniy glagola iakutskogo iazyka [Linguistic annotation of the verb moods of the Yakut language]. In *Vestnik SVFU* [*North-Eastern Federal University Newsletter*], 3, 108-120.

Zheltov, P.V. (2015). Morphological markup system for the national body of the Chuvash language. In *Proceedings of the International Conference "Turkic Languages Processing: TurkLang-2015",* Kazan, Academy of Sciences of the Republic of Tatarstan Press, 328-330.

Лингвистическое аннотирование грамматических категорий языка саха: имя существительное

Г.Г. Торотоев, С.Г. Торотоева
Северо-Восточный федеральный университет им. М.К. Аммосова
Россия, 677000, Якутск, ул. Белинского, 58

Статья посвящена работе по созданию инструментария для лингвистического аннотирования грамматических категорий языка саха. Базируясь на Лейпцигских правилах глоссирования, описываем основные словоизменительные характеристики имени существительного в якутском языке (число, персональность, посессивность, падежная система). В результате научно-изыскательских работ (2014-2018) разработана система тэгов, отображающая весь словоизменительный потенциал имени существительного в якутском языке, включающий в своем арсенале 247 морфологических показателей. Разрабатываемая тюркологами унифицированная система морфологической разметки тюркских языков далеко не совершенна, существуют различные трактовки по части отображения и интерпретации грамматических категорий в разных тюркских языках. Несмотря на это в статье обобщены конструктивные идеи коллег по данной проблематике.

Ключевые слова: лингвистическое аннотирование, грамматические категории, язык саха, имя существительное, число, посессивность, простое склонение, притяжательное склонение, диминутив, тэги.

Научная специальность: 10.00.00 – филологические науки.