УДК 81’366

МОРФОЛОГИЧЕСКИЕ РАЗМЕТКИ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА

Т.С. Садыков, Б.Ш. Шаршембаев, Б.О. Көчкөнбаева

Идентификация категорий частей речи и их распределение в текстах является одной из сложнейших проблем тюркологии. В отличие от флективных языков в тюркских языках слова и словоформы чаще подвергаются явлениям конверсии и омонимии. Для разметки морфологических категорий слов и словоформ в корпусе, а также для снятия текстовой омонимии предлагается система морфологической разметки, совместимой с унифицированной системой для национального корпуса тюркских текстов.

Ключевые слова: части речи; морфологические категории; система морфологической разметки; национальный корпус.

MORPHOLOGICAL TAGGING FOR THE NATIONAL CORPUS

T.S. Sadykov, B.S. Sharshembaev, B.O. Kochkanbaeva

The identification of parts of speech and their distribution in the texts is one of the most complicated problems of Turkology. In contrast of inflectional languages in Turkic languages words and word forms undergo phenomenon of conversion and homonymy. For marking of morphological categories of words and word forms in the corpus, and also for theremoval of a text homonymy, the system of morphological tagging compatible to the unified system for the national corpus of Turkic texts is offered.

Keywords: parts of speech; morphological categories; system of the morphological tagging; national corpus.

Учурдагы тил теориясындагы парадигматикалык жылыштарды байкап багып, жакындан аңдап туюп, морфемика, сөз жасоо, социолингвистика, маданий тилтааным, когнитивдик лингвистика багыттарында топтолгон зор илимий тажрыбасын колдонуу аркылуу орто мектептер үчүн жетик окуу китептерин жазган, жождордо тил илиминин өзөк маселелери боюнча узак жылдар бою дарс окуп келген, республикабызда бири мамлекеттик, экинчиси расми тил мартабасын алган кыргыз жана орус тилдеринин колдонушу менен өз ара таасирин “тилдер жана маданияттар диалогу” катары карап, коомдук мамиле, тилдик аң-сезим, билингвизм, тилдик тулга, таттуу мамиледеги тилдик коом сыяктуу өңүттөрдөн иликтеп, булардын табиятын илимий жактан териштирген проф. М.Тагаевдин салган чыйырын улай бул макалада кыргыз тилинин улуттук корпусун түзүүдө морфологиялык энтектер маселеси талкуга алынмакчы.

Морфологиялык энтектер (morphological tegs) – сөз, сөз формасы, уңгу, мүчө сыяктуу тексте колдонулган өзөк бирдиктерди талдап, аларды морфологиялык категориялардын шарттуу энтек (эн + teg) белгилери менен эн-белгилөө системасы. Кыргыз текстинин улуттук корпусун морфологиялык энтектер менен белгилеп чыгуу үчүн Казан шаарында түрк тилдери боюнча улуттук корпусту түзүү ассоциясы тарабынан кабыл алынган белгилердин стандарты жетекчиликке алынат [1, с. 140–147; 2; 3].

Энтектер системасы менен жабдылып, адегенде морфологиялык анализден, андан соң синтаксистик, семантикалык, прагматикалык талдоодон өткөрүлгөн текстердин жыйындысы компьютер түшүнө турган жандуу корпуска айланат. Тактап айтканда, мындай улуттук корпус компьютер үчүн формалдаштырылган, автоматтап иштетүүгө даяр тексттердин корпусу болсо, колдонуучу үчүн лингвистикалык, когнитивдик, этнологиялык, лингвокультурологиялык өңүттөрдөнолуттуу маалыматтарды алуунун *улуттук корпусу же билик базасы (knowledge base, intelligent database, база знаний)* болот [4].

Тектеш түрк тилдеринде бул багытта унификациялоо иши аткарылбагандыктан, бирдей морфологиялык категориялар демейдеар башкача белгиленип келүүдө. Улуттук корпусту түзүү менен алектенгентилчи-окумуштуулар, компьютердик лингвистер, инженер-программистертексти

эн-белгилөөдө колдонулган энтектектерди унфикациялоо керектигин аңдап түшүнүшүүдө. Ошол эле маалда окшош кубулуштарды бир стандартта бирдей белгилөө жагынан энтектектерди унфикациялоо жеке эле тектеш тилдер үчүн эмес, тектеш эмес тилдер үчүн да зарыл экендиги ортого чыгууда.

Ошентип, 2014-жылы Казан шаарында Татарстан Республикасы Илимдер академиясы менен Казан федералдык университетинин демилгеси менен уюшулган Түрк тилдеринин улуттук корпустарын түзүү ассоциясы тарабынан кабыл алынган энтектер стандарты типологдордун Лейпциг стандарты менен шайкеш келет да, тексте колдонулган жана тиги же бу сөз түркүмүнө тиешелүү болгон ар сөзгө эн-белги салынып, төмөнкү энтектер аркылуу белгиленет [3].

Энтектер tags	**Аталышы fullterm**	**Сөз түркүмү parts of speech**
N	noun	зататооч
ADJ	adjective	сын атооч
V	verb	этиш
ADV	adverb	тактооч
NUM	numeral	сан атооч
PN	pronoun	ататооч
CNJ	conjunction	байламта
POST	postposition	жандооч
PART	particle	бөлүкчө
INTRJ	interjection	сырдыксөз
MOD	modalword	модалдык сөз
IMIT	imitativeword	тууранды сөз

Ошол эле маалда тексте колдонулган ар бир сөз жана сөз формасын грамматикалык категориялар боюнча белгилөө үчүн да төмөнкү энтектер системасы сунушталат.

Сан категориясы – Number: 1. Жекелик сан – singular 2. Көптүк сан – plural

Энтектери: 1. SG <=> Ø

2. PL <=> ЛАр

(-лар -дар -тар/-лер -дер –тер/-лор -дор -тор/-лөр -дөр -төр).

Таандык категория – Possessive

Жекелик сан – singular:

1. Таандык жекелик сан 1-жак – 1st person singular possessive (‘my’),
2. Таандык жекелик сан 2-жак – 2nd person singular possessive (‘your’),
3. Таандык жекелик сан 2-жак сылык – 2nd person sing.poss. formal (‘your’),
4. Таандык жекелик сан 3-жак – 3nd person singular possessive (‘his/her/its’),

Көптүк сан – plural:

5. Таандык көптүк сан 1-жак – 1st person plural possessive (‘our’),
6. Таандык көптүк сан 2-жак – 2nd person plural possessive (‘your’),
7. Таандык көптүк сан 2-жак сылык – 2nd person pl.poss. formal (‘your’),
8. Таандык жекелик сан 3-жак – 3nd person plural possessive (‘their’),

Энтектери:

1. **POSS_1SG <=> [Ы]м:** -ым -им -ум -үм - м
2. **POSS_2SG <=> [Ы]ң:** -ың -иң -уң -үң - -ң;
3. **POSS_2SGF <=> [Ы]ң[Ы]з:** -ыңыз -иңиз -уңуз -үңүз – -ңыз -ңиз -ңуз -ңүз;
4. **POSS_3SG <=> [с]Ы[н]:** -ы -и -у -ү -ын -ин -ун –үн; -сы -си -су -сү -сын -син -сун -сүн;
5. **POSS_1PL <=> [Ы]б[Ы]з:** -ыбыз -ибиз -убуз –үбүз/-быз -биз -буз -бүз;
6. **POSS_2PL <=> [Ы]ң[А]р:** -ыңар -иңер -уңар -үңөр /-ңар -ңер -ңар -ңөр;
7. **POSS_2PLF <=> [Ы]ң[Ы]зд[А]р:** -ыңыздар-иңиздер -уңуздар -үңүздөр /-ңыздар -ңиздер -ңуздар -ңүздөр;
8. **POSS_3PL <=> [с]Ы[н]:** -ы -и -у -ү /-ын -ин -ун –үн/-сы -си -су -сү / -сын -син -сун -сүн;

Жөндөмө категориясы – Noun Cases

1. Атооч жөндөмө – nominative, 2. Илик жөндөмө – genitive, 3. Барыш жөндөмө – dative, 4. Табыш жөндөмө – accusative, 5. Жатыш жөндөмө – locateve, 6. Чыгыш жөндөмө – ablative.

Энтектери: 1. NOM<=>Ø – Ø

2. **GEN <=> [н]Ын** – -нын -нин -нун -нүн /-дын -дин -дун -дүн /-тын -тин -тун –түн/-ын -ин -ун -үн;
3. **DAT<=>[Г]А:** -га -ге -го –гө/-ка -ке- ко –кө/-а -е- о -ө;
4. **ACC<=> [н][Ы]:** -ны -ни -ну –нү/-ды -ди -ду –тү/-ты -ти -ту –тү/-ы -и -у -ү;
5. **LOC<=>ДА:** -да -де -до –дө/-та -те -то -тө;
6. **ABL<=> [Д]Ан**: -дан -ден -дон –дөн/-тан -тен -тон –төн/-ан -ен -он -өн;

Жак категориясы – Personal

Жекелик сан – singular:

1. 1-жак жекелик сан – 1st person singular,
2. 2-жак жекелик сан – 2nd person singular,
3. 2-жак жекелик сан сылык – 2nd person singular formal,
4. 3-жак жекелик сан – 3nd person singular,

Көптүк сан – plural:

5. 1-жак көптүк сан – 1st person plural,
6. 2-жак көптүк сан – 2nd person plural,

7. 2-жак көптүк сан сылык – 2nd person plural formal,
8. 3-жак көптүк сан – 3nd person plural,

Энтектери: 1. 1SG<=>м[Ы]н – -мын -мин -мун -мүн; **2. 2SG <=> с[Ы]ң** – -сың -сиң -суң -сүн; **3. 2SGF <=> с[Ы]з** – -сыз -сиз -суз -сүз; **4. 3SG <=>Ø Ø; 5. 1PL <=> б[Ы]з** – -быз -биз -буз -бүз; **6. 2PL <=> с[Ы]ң[А]р** – -сыңар -сиңер -суңар -сүңөр; **7. 2PLF <=> с[Ы]зд[А]р** – -сыздар-сиздер -суздар -сүздөр; **8. 3PL <=> [с]Ы[н]** – Ø;

Сын атооч – adjective: салыштырма даража – comparative

Энтеги: COMP <=> [Ы]рААК – -ыраак -ирээк -ураак –үрөөк/-раак -рээк -раак -рөөк;

Сан атооч – Numeral

1. Иреттик сан – ordinal numeral 2. Жамдама сан – collective numeral, 3. Чамалама сан1 – approximatenumeral 1, 4. Чамалама сан 2 – approximatenumeral 2, 5. Чамалама сан 3 – approximatenumeral 3,

Энтектери:

1. **NUM_ORD <=> [Ы]нчЫ:** -ынчы -инчи -унчу -үнчү /-нчы -нчи -нчу -нчү;
2. **NUM_COLL <=> ОО[н]:** -оо –өө/-оон -өөн;
3. **NUM_APPR1 <=> чА:** -ча -че -чо -чө;
4. **NUM_APPR2 <=> ДАй:** -дай -дей -дой -дөй /-тай -тей -той -төй;
5. **NUM_APPR3 <=> ДАгАн:** -даган -деген -догон -дөгөн /-таган -теген -тогон -төгөн.

Этиш – Verb: Мамиле категориясы – Voices

1. Негизги мамиле – active, 2. Туюк мамиле – passive, 3. Өздүк мамиле – reflexive, 4. Аркылуу мамиле – causative, 5. Кош мамиле – reciprocal.

Энтектери:

1. ACT<=> Ø – Ø 2. ACT <=> [Ы]л||н: -ыл -ил -ул –үл/-л;/-ын -ин -ун –үн/-л; **3. REFL<=>[Ы]н:** -ын -ин -ун –үн/-л; **4. CAUS <=>Д[ЫР] 5. RECP <=> [Ы]ш:** -ыш -иш -уш –үш/-ш;

Буйрук ыңгай – Imperatives

1. 1-жак жекеликсан – Hortative: 1st person singular – ‘let me’,
2. 1-жак көптүк сан – Hortative: 1st person plural – ‘let’s’,
3. 2-жак жекеликсан сылык – Imperative: 2nd person singular,
4. 2-жак көптүк сан сылык – Imperative: 2nd person plural,
5. 2-жак жекеликсан сылык – Imperative: 2nd person singular formal,
6. 2-жак көптүк сан сылык – Imperative: 2nd person plural formal,
7. 3-жак жекеликсан – Jussive: 3rd person singular – ‘let him/her/it’,
8. 3-жак көптүк сан – Jussive: 3rd person plural – ‘let them,
9. Өтүнүч сылык – precative (‘please’).

Энтектери:

1. HOR_SG <=> [А]йЫн: -айын -ейин -ойун -өйүн, -йын -йин -йун –йүн;

2. HOR_PL <=> [А||й]лЫ[к]: -алык -елик -олук -өлүк/-йлык -йлик -йлук -йлүк;/-алы -ели -олу -өлү/-йлы -йли -йлу -йлү;

3. IMP_SG <=> ГЫн: -гын -гин -гун -гүн/-кын -кин -кун -күн;

4. IMP_ PL <=> ГЫлА: -гыла -гила -гула -гүла/-кыла -кила -кула -күла;

5. IMP_SGF <=> [Ы]ңЫз: -ыбыз -ибиз -убуз -үбүз/-быз -биз -буз -бүз;

6. IMP_PLF <=> [Ы]ңЫздАр: -ыңыздар-иңиздер -уңуздар -үңүздөр /-ңыздар -ңиздер -ңуздар -ңүздөр;

7. JUS_SG <=>сЫн: -сын -син -сун -сүн;

8. JUS_PL <=> [Ыш]сЫн: -ышсын -ишсин -ушсун -үшсүн/-сын -син -сун -сүн;

9. PREC_1 <=>чЫ: -чы -чи -чу -чү;

Чак категориясы – Verb tenses

1. Учур чак – present, 2. Айкыр өткөн чак – past definite, 3. Жалпы өткөн чак – past indefinite, 4. Капыскы өткөн чак – past evidentiality, 5. Адат өткөн чак – past iterative, 6. Айкын келер чак – future definite, 7. Арсар келер чак – future indefinite, 8. Арсар терс келер чак – future indefinitenegative.

Энтектери:

1. PRES<=> [А||й]: -а -е -о –ө/-й; **2. PST_DEF <=> ДЫ:** -ды -ди -ду -дү /-ты -ти -ту -тү; **3. PST_INDF<=> ГА[н]:** -ган -ген -гон -гөн/-кан -кен -кон -көн/-га -ге -го -гө/-ка -ке -ко -кө; **4. PST_EVID <=> чУ:** -чу -чү; **5. PST_ITER <=>[Ы]п[тЫр]:** -ыптыр -иптир -уптур -үптүр, -птыр -птир -птур -птүр; -ып -ип -уп -үп - -п; **6. FUT_DEF <=> [А||й]:** -а -е -о -ө - -й; **7. FUT_INDF<=>[А]р:** -ар -ер -ор -өр - -р; **8. FUT_INDF_NEG <=> БАс:** -бас -бес -бос -бөс/-пас -пес -пос -пөс;

Аспект – aspect: 1. Тануу – negative, 2. Суроо – interrogative.

Энтектери:

1. NEG <=> БА: -ба -бе -бо -бө/-па -пе -по -пө;

2. INT<=> БЫ: -бы -би -бу -бү/-пы пи -пу -пү.

Атоочтуктар – Participles: 1. Учур чак атоочтук – present participle 2. Өткөн чак атоочтук – pastparticiple, 3. Келер чак атоочтук – future participle, 4. Келер чак терс атоочтук – future participlenegative.

Энтектери: 1. PCP_PR <=> [УУ]чУ: -уучу -үүчү/-чу -чү;

2. PCP_PS <=> ГАн: -ган -ген -гон -гөн/-кан -кен -кон -көн;

3. PCP_FUT_DEF<=> [А]р: -ар -ер -ор -өр - -р;

4. PCP_FUT_ NEG <=> БАс: -бас -бес -бос -бөс/-пас -пес -пос -пөс;

Чакчылдар – Converbs: 1. Коштоочу чакчыл - Adverbial verb (accompanist), 2. Созулма чакчыл – Adverbial verb (continuing), 3. Максат чакчыл – Adverbial verb (Intentional), 4. Тангыч чакчыл – Adverbial verb (negative form), 5. Удаалаш чакчыл – Adverbial verb (successive meaning), 6. Чектеме чакчыл - Adverbial verb (limiting).

Энтектери: 1. ADVV_ACC <=> [Ы]п: -ып -ип -уп -үп - -п; **2. ADVV_CONT <=> [А||й]:** -а -е -о –ө - -й; **3. ADVV_INT <=> ГАнЫ:** -ганы -гени -гону -гөнү/-каны -кени -кону -көнү; **4. ADVV_NEG <=> мАйЫн[чА]:** -майынча -мейинче -мойунча-мөйүнчө;/-майын -мейин -мойун -мөйүн; **5. ADVV_SUC <=> ГЫчА:** -гыча -гиче -гуча -гүчө/-кыча -киче -куча -күчө; **6. ADVV_SUC <=> ГАнчА:** -ганча -генче -гончо -гөнчө/-канча -кенче -кончо -көнчө.

Кыймыл атооч – Verbal nouns (masdars): 1. Кыймыл атооч -оо – infinitive 1, 2. Кыймыл атооч -уу – infinitive 2, 3. Кыймыл атооч -ыш – infinitive 3, 4. Кыймыл атооч -мак – infinitive 4, 5. Кыймыл атооч -гы – infinitive 5.

Энтектери:

1. INF_1 <=> ОО - -оо -өө; **2. INF_2 <=> УУ** - -уу -үү; **3. INF_3 <=> [Ы]ш** - -ыш -иш -уш -үш - -ш; **4. INF_4 <=> мАГ** – -мак -мек -мок -мөк/-маг -мег -мог -мөг; **5. INF_5 <=> ГЫ** – -гы -ги -гу -гү/-кы -ки -ку –кү.

Модалдык формалар – Modal forms:

1. Шарттуу модалдык – conditional, 2. Ниет модалдык – desiderative (intention), 3. Тилек модалдык – optative1, 4. Тилек модалдык – optative 2, 5. Кооп модалдык – premonitive (warning).

Энтектери:

1. COND <=> сА: -са -се -со -сө; **2. DESIDE <=> мАк[чЫ]:** -макчы -мекчи -мокчу –мөкчү/-мак -мек -мок -мөк; **3. OPT <=> ГЫ+POSS** келет||келди: -гы -ги -гу –гү/-кы -ки -ку -кү; **4. OPT <=> ГАй эле+PERS:** -гай -гей -гой –гөй/-кай -кей -кой -көй; **5. PREM <=> БАГАй эле+PERS:** -багай -бегей -богой –бөгөй/-пагай -пегей -погой –пөгөй.

Литература:

1. *Садыков Т.* Система морфологической разметки для корпуса кыргызских текстов / Т. Садыков, Б. Шаршембаев // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL – 2014. Казань: Изд-во Фан АН РТ, 2014. С. 140–147.
2. URL: http://ips.antat.ru/page.php;http://www.eva.mpg.de/lingua/resources/ glossingrules.php
3. URL: http://www.eva. mpg.de/lingua/resources/glossingrules.php
4. *Шубина О.Ю.* Понятие грамматического морфологического значения в языке / О.Ю. Шубина // Вестник КРСУ. 2011. № 1. С. 93–97.