УЎК (УДК, UDC): 81'322.2
DOI: 10.36078/1570167968

ЎЗБЕК ТИЛИДА *WORDNET* ЯРАТИШ МАСАЛАЛАРИГА ДОИР

Нилуфар Зайнобиддин қизи АБДУРАҲМОНОВА
Филология фанлари бўйича фалсафа доктори (PhD)
Алишер Навоий номидаги Тошкент давлат
ўзбек тили ва адабиёти университети
Тошкент, Ўзбекистон
abdurahmonova.1987@mail.ru

Муҳаммад Рустам ўғли ҲАЙДАРОВ
Талаба
Алишер Навоий номидаги Тошкент давлат
ўзбек тили ва адабиёти университети
Тошкент, Ўзбекистон
haydarovmuhammad777@gmail.com

Аннотация

Компьютер лингвистикаси соҳаси учун лексикографик тадқиқотлар муҳим аҳамиятга эга. Айниқса, ўзбек тилининг миллий корпусини яратиш ҳамда электрон луғатларни такомиллаштириш бу соҳадаги кўплаб йўналишларни ривожлантиришга омил бўлиб хизмат қилади. Ўзбек тилида WordNet яратишда олдиндан яратилган тайёр лингвистик моделлар асос бўлиб хизмат қилади. Ушбу мақолада ўзбек тилининг UzNet луғатини яратиш учун лингвистик тадқиқотларга доир илмий мулоҳазалар баён қилинган.

Калит сўзлар: Wordnet; Synset; компьютер лексокографияси; синоним; антоним.

О ЗАДАЧАХ СОЗДАНИЯ *WORDNET* НА УЗБЕКСКОМ ЯЗЫКЕ

Нилуфар Зайнобиддин кизи АБДУРАХМОНОВА
Доктор PhD
Ташкентский государственный университет узбекского языка
и литературы им. Алишера Навои
Ташкент, Узбекистан
abdurahmonova.1987@mail.ru
Мухаммад Рустам угли ХАЙДАРОВ
Студент
Ташкентский государственный университет

узбекского языка и литературы
им. Алишера Навои
Ташкент, Узбекистан
haydarovmuhammad777@gmail.com

Аннотация

Иследования в сфере лексикографии важны для компьютерной лингвистики. Особенно в области составления электронного словаря и создания национального корпуса узбекского языка может быть основным фактором развития нескольких направлений в этой сфере. Прежние лингвистические модели являются основой для WordNet в узбекском языке. В этой работе указаны научные основания для лингвистических иследований для создания UzNet узбекского языка.

Ключевые слова: WordNet; Synset; компьютерная лексикография; синоним; антоним.

ON THE TASKS OF CREATING A *WORDNET* IN THE UZBEK LANGUAGE

Nilufar Zaynobiddin kizi ABDURAKHMONOVA
PhD
Alisher Navoi Tashkent State
Uzbek Language and Literature University
after Alisher Navoi
Tashkent, Uzbekistan
abdurahmonova.1987@mail.ru

Muhammad Rustam ugli KHAYDAROV
Student
Alisher Navoi Tashkent State Uzbek
Language and Literature University after Alisher Navoi
Tashkent, Uzbekistan
haydarovmuhammad777@gmail.com

Abstract

The research in lexicography is important for computational linguistics. Particularly, creation electronic dictionary and building national corpus of the Uzbek language could be the main factor for the development of a number of directions in this sphere. Previous linguistic models are groundwork for WordNet in the Uzbek language. The paper deals with scientific ideas about linguistic investigations for creation UzNet of the Uzbek language.

Keywords: WordNet; Synset; computational lexicography; synonym; antonym.

1985 йили Принстон университетининг бир неча лингвист ва психологлари WordNet номли лексик маълумотлар базасини яратиш лойиҳаси бўйича тадқиқот олиб бордилар. Ўша даврда у “лексик онтология” деб ҳам номланган. Психолингвист Жорж Миллер инсоннинг семантик хотирасини тушунишга ҳаракат қиладиган сунъий интеллект устида олиб

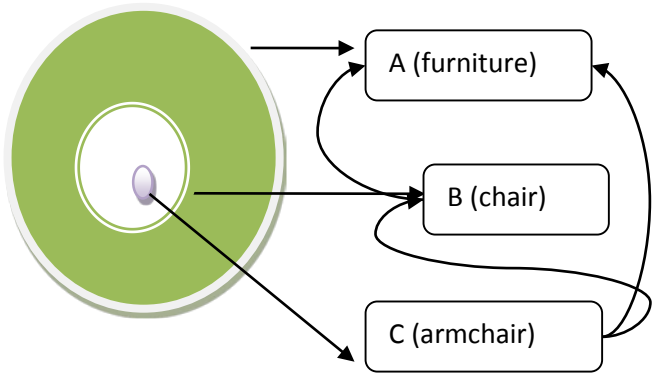
борган тажрибаларидан илҳомланган. Унга кўра WordNet тахминан 95600 сўзшаклларининг (51,500 сўз ва 44,100 сўз бирикмалари) 70,100та маъноси, синонимлар билан берилган. Шулардан тахминан 57000таси отга тегишли сўзшакллари ва улардан 47000тасининг маънолари киритилган. Мақолалар миқдори тахминий айтилишининг сабаби шундаки, онлайн тизимда бу кўрсаткич доимий равишда ўсиб, ўзгариб туради. WordNet билан бошқа анъанавий луғатларнинг фарқли жиҳати — WordNet фақат от, феъл, сифат ва равишдан ташкил топади. WordNet (Феллбаум, 1998) психолингвистик тадқиқотларга доир лексик маълумотлар базасидир. Принстон университетида тузилган бу тизим тил муҳандислиги тадқиқотларида фойдаланилади. WordNet от ва феъл туркумидаги сўзларнинг синоними ва гипонимлари ҳамда сифатлар учун антонимлар асосида тузилган. WordNet феъллар учун тропонимлар ва отлар учун гипонимлар терминини фарқлайди. Ҳар бир сўзнинг маъноси synsetда (синимлар тизимида) берилиб, иерархик ва ўзаро семантик алоқалари ифодаланган.

1.	Nature, wild, natural state, state of nature- (a wild primitive state untouched by civilization; “he lived in the wild; “they tried to preserve nature as they found it”) =>state-(the way something is with respect to its main attributes; “the current state of knowledge; “his state of health; “in a weak financial state”)
2.	Natural phenomenon, nature-(all non-artificial phenomenon) => phenomenon-(any state or proccess known through trough the senses rather than by intuition or reasoning)

Эндиликда WordNetнинг турли тилларга доир ресурси яратилган ва у тобора хусусий базадан глобал базага қараб кенгайиб бормоқда (жумладан, албан тили учун AlbaNet, дания тили учун DaNet, болгар тили учун BulNet, поляк тили учун PolNet ва бошқалар).

WordNet инглиз тилининг лексик маълумотлар базаси ҳисобланади. Ундаги от, феъл, сифат ва равиш сўз туркумлари муайян концептни ифодаловчи синонимлар гуруҳига (Synset) бирлашган. Synsetсетлар концептуал-семантик ва лексик муносабатларга кўра ўзаро боғланган. WordNetдан фойдаланиш ва уни ўрнатиб олиш қулай. Унинг тузилиши компьютер лингвистикаси ва табиий тилни қайта ишлаш учун фойдали. WordNet айрим жиҳатлари билан тезаурусга ўхшайди, бироқ айрим фарқли

жиҳатлари ҳам кузатилади. WordNetнинг ички боғланиши сўзшакли (ҳарфлар қатори) билан эмас, сўзларнинг маънолари билан юзага келган. Натижада сўзлар семантик кўп маъноли бўлган бошқа сўзларга яқин бўлган вариант билан топилади. WordNetда семантик муносабатларидаги сўзлар таснифланган бўлиб, тезаурусда муайян сўзга яқин бўлган сўзлар учрамайди (6). WordNetдаги асосий муносабат синонимияга қурилган: ***катта-улкан, йиғламоқ-ҳўнграмоқ*** каби. Бир хил концептга бирлаштирилган ва контекстдаги ўзаро ўзгарадиган синонимлар тартибланмаган гуруҳларда умумлаштирилган. WordNetдаги 117 000 Synset кичик ҳажмли концептуал муносабатларга кўра бошқа Synsetларга боғланган. Synsetда қисқа изоҳлар (глос) берилган бўлиб, кўп ҳолларда бир ёки ундан ортиқ Synsetдаги бирликларни қўлланилиши акс этган. WordNetдаги ҳар бир сўзнинг маъновий шакли битта бўлади. WordNetдаги сўзлар субординацион, яъни бутун — қисм муносабатига қурилган. Сўзлар иерархик равишда юқоридан қуйи томон (giperonim, giponim) умумий {furniture, piece_of_furniture} Synsetдан махсус Synsetга қадар {bed} ва {bunkbed} бирлашади. Шундай қилиб WordNetда *мебел* категорияси *ётоқ* ва *икки кишилик кровать* сўзларига боғланади. Барча от сўз туркумига тегишли сўзлар ўзагига қараб ўсиб боради. Гипоним сўзлар маъно жиҳатидан бошқа сўзларга ҳам алоқадор бўлиши мумкин: *агар кресло курсининг тури бўлса ва курси мебелнинг тури бўлса, у ҳолда курси ҳам кресло ҳам мебелнинг тури саналади.* Буни қуйидаги ифода билан бериш мумкин (чизма):



Чизма

Муҳим жиҳатлардан бири — синонимия ва полисемия таржимада бир қанча мураккабликларни юзага келтиради. Масалан, бир сўзнинг бир нечта маънолари ва уларнинг синонимлари бўлиши мумкин. Унда синонимия сўзшакллар ўртасидаги лексик муносабатлар ҳисобланиб, синонимияда сўзларнинг ўртасидаги асосий фарқи учун {} белгиси, бошқа қўшимча лексик алоқалар учун [] белгилари билан белгиланади.

WordNet тизимида инглиз тилидаги барча сўз туркумларининг семантик маънолари ифодаланган ва у семантик алоқалар орқали ташкиллаштирилади. От сўз туркуми мисолида ушбу луғатнинг хусусиятлари таҳлилга тортилар экан, унинг ҳажми 80000 дан зиёд от сўз туркумига тўғри келиши ва бу кўрсаткичда муайян лексемани нутқий бирлик сифатида воқеалинишида сўз бирикмаси ҳолатида қўлланиш имконияти қамраб олинган. WordNet компьютер лексикографиясида эришилган инновацион ютуқлардан биридир. Чунки унга илова этилган сўзликлар машина ўқиши учун мумкин бўлган ҳолатга мослаштирилган. Одатий луғатларда киритилган сўзликларнинг талаффузи, грамматик шаклланиши, ясалиши, этимологияси, изоҳи ҳамда синоним, антоним каби яна бир нечта лингвистик хусусиятлар сингдирилади. Бундай белгиларнинг машина ўқиши мумкин бўлган имконияти мавжуд бўлмаганлиги боис аксарият жиҳатлари тушириб қолдирилади, хусусан, WordNet юқорида қайд этилган талаффуз, этимологик тавсиф ва шунга ўхшаш маълумотларни ўз ичига қамраб олмайди. Ушбу луғатдан ўрин олган мақолаларнинг қўлланилишида ифода маъноларининг семантик хусусиятларини очиб бериш билан боғлиқ афзалликларни яратиш, қолаверса, тўғри таржимага эришишда сўз майдонларини мантиқий лойиҳалаш бирламчи мезон ҳисобланади. Шу жиҳатдан WordNet сўзларнинг семантик алоқаларини чуқур ўрганишда синонимия ҳодисаси муҳим деб қаралади. WordNetда синонимлар муайян блокларга ажратилиб, семантик жиҳатдан таснифланган. Спарск Жонес (1964, 1986) ўзининг семантик тасниф назариясига асос солган тадқиқотчидир. У ўзининг бу борадаги изланишларида матндан олинган муайян сўз шакли бошқа сўз шакллари билан бирга бўлиши мумкин бўлган барча синонимлар тизимини яратади. Масалан, у инглиз тилидаги қурол-яроғ билан боғлиқ бўлган тушунча ***pellet*** ҳамда ***injection*** (инекция) сўзларининг контекстдаги маънодоши ***shot*** лексемаси билан боғлайди, бу эса шартли равишда муайян ҳолат учун синонимик қаторни ҳосил қилади. Синонимлар тизими (Synset) ўзаро семантик боғланса-да, ҳар икки тизимнинг структураси ҳар хил типга тегишли бўлади: {*shot, pellet*} ва {*shot, injection*}. Бу икки тўплам маъно жиҳатидан ўзаро ҳеч қандай алоқага киришмайди, яъни уларни фақат *shot* лексемасигина боғлаб туради.

Аксарият Synsetда одатий луғатлардаги изоҳлардан фойдаланилади. Полисемантик сўзларда бир неча кўчма маънолар ифодаланган глоссалар

мавжуд бўлса, Synsetда фақат ягона глоссема мавжуд бўлади. WordNetдаги Synset орқали сўзликларнинг лексикаллашув концепти шаклланади.

Синонимлар ўртасида семантик боғланиш бўлса-да, семантик боғланиш отларнинг лексикаллашув концепти ўртасидаги боғланишнинг муҳим тармоғи саналади. Булар субординацияларнинг (гуруҳ ёки муайян таснифларнинг) боғланиши бўлиб, унга кирувчи элементлар гипонимлар деб аталади. Масалан, *бургут қуш* сўзининг гипоними бўлади, *қуш* эса *бургут* сўзининг гипероними бўлади. Бундай семантик боғланиш лексик иерархияни ҳосил қилади. Одатий луғатларда ҳам отлар ўртасидаги гипонимик боғланиш ҳақидаги маълумотлар берилади (Амслер, 1980).

Лексик иерархия қуйидагича тармоқланади: {robin, redbreast} @.–> {bird} @□> {animal, animate_being} @□> {organism, life_form, living_thing}.

@.–>белги муайян сўзни бошқа сўзга қараб хусусийлашиб бориши, яъни генерилазациялашувини билдиради. Сс @–>ушбу белги от сўз туркумидан ташкил топган Synsetни бошқа Сг ўтишини ифодалайди. Яъни у «тури» деган маънони ҳам ифодалиши мумкин. Жумладан, Synset ўртасида доимий инверсия ҳодисаси содир бўлади: Сг ~ –> Сс. бошқача қилиб айтганда, Сс Сгнинг гипероними, Сг эса Сснинг гипоними ҳисобланади. ~->бу белги гиперонимдан гипонимга қараб маъно кенгайишини ифодалайди. Нафақат {bird} Synset гипероним билан боғланган бўлади ва балки бунинг таркибига барча қуш турлари ҳам киритилади. Бундай боғланиш ҳақида маълумотларни бериш ҳамда уларни тузиш бўйича қўйилган талаблар одатий луғатларда ҳам мавжуд, бироқ мақолаларни топишни янада осонлаштириш зарур. Луғатшунослар сўзликларни киритишда у ҳақидаги маълумотни айлана ҳолатда эмас, балки тармоқли ҳолатда берса мақсадга мувофиқ, деб ҳисоблайдилар. Тармоқланиш компьютер технологияларида ҳам самарали усуллардан биридир. Бу метод маълумотлар базасида умумий бўлиб, мавжуд бирликлар ҳақидаги махсус жойга ID рақами билан ўзаро боғлиқ маълумотларни бириктириш имконини яратади. Иерархияллашув катта ҳажмли маълумотлар базасини шакллантиришда компютер дастурчилари томонидан кенг қўлланилади (Touretzky, 1986). Чунки тармоқланиш бирликларга ажратилаётган жойни тежашда самарали ҳисобланади.

Компьютер мутахассислари бу усулни “ворис тизими” деб номлашади, чунки мерос кейинги авлодга тармоқланган ҳолда узатилади. Лексик структура чизиқли ҳолатда акс этади: — *Агар ой ернинг юлдоши бўлса, демак ой сайёра; агар у сайёра бўлса, қуёш тизимининг таркиби; агар ой қуёш*

тизимининг тизими бўлса, демак бир хил динамик ҳолатда ҳаракатланувчи. Синоним, антоним ва гипонимлар ўзаро семантик алоқаси мавжуд тушунчалардир. Шунингдек, WordNetда **меронимлар** (бутун-қисм муносабатлари) ва антонимлар ҳам ўрин топган. Масалан,{ғилдирак}сўзи {велосипед} сўзининг мероними ҳисобланади.

Wm #p —► Wh — бу ерда Wm Whнинг нинг таркибий қисми эканлигини биладиради;

Wm #m —> Wh — Wm Whнинг аъзоси эканлигини билдиради;

Wm #s —► Wh — Wm Whнинг таркиби эканлигини билдиради;

#p — WordNetдаги сўзларнинг қисм тўплам алоқасини билдиради.

Лексик муносабатларнинг муҳим синфини сўзшакллари ва морфологик алоқалари ташкил қилади. Зотан, WordNetга сўзнинг муайян грамматик категорияси билан биргаликда сўз киритилганда унинг бошланғич шаклини топиш зарур: *Books=>book, geese=> goose; go-went-gone-going* каби. Иерарархик тарзда сўзларнинг таснифланиши муайян чегараларга эга. Масалан, мероним бутун қисм муносабати фақат от сўз туркуми доирасида бўлса, антоним ҳам шундай ҳолатда бўлади. Шу жиҳатдан сўзларнинг уч жиҳати инобатга олинади: атрибутлари, қисмлари (мероним), функциялари. Шу жиҳатдан WordNet турли семантик компонентларни ўз ичига олган йигирма бешта файлдан иборат. WordNetдаги асосий муносабат бир хил сўз туркуми (СТ) орасида ҳосил бўлган. WordNet сўз туркумларидан иборат ички гуруҳга ва айримлари кесишган СТ (ПОС) бўлинган. Ушбу кесишувчи СТ муносабатлари ўзакдош бўлган морфосемантик алоқадаги тизимдан иборат: *observe* (verb), *observent* (adjective), *observation*, *observatory* (noun). От-феъл жуфтликлари от билан феълнинг семантик муносабатида ўзаро жуфтлашади: {sleeper, sleeping_car} {sleep} учун МАКОН ва {painter}{paint}нинг АГЕНТи, шунда {painting, picture} НАТИЖА бўлади.

Synset структураси

Семантик муносабат	**Таърифи**	**Сўз туркуми**				**Мисол**
		От	Феъл	Сифат	Равиш	
синоним	Маъно жиҳатдан бир-бирига яқин сўзлар	+	+	+	+	{sofa, couch, lounge}=> giperonim {seat}
антоним	Маъноси бир-	+	+	+	+	{love} ⇔{hate,

	бирига зид бўлган сўзлар					detest}
гипероним	Тур ва тасниф муносабатларини жамловчи концепт	+	+			{feline, felid}giperonim<= {cat, true cat}
гипоним	Муайян турга тегишли сўзлар	+	+			{wild cat} giponim<={cat, true cat}
мероним	Моҳиятан бирор концептга тегишли сўзлар	+				{snowflake, flake} моҳияти<={snow}

WordNetда феъл сўз туркумига кирувчи **тропонимлар** киритилган. Унга кўра у ёки бу феъл бошқа муайян шаклига кўра боғланади. Масалан, *вайсамоқ гапирмоқ* фълининг тропоними ҳисобланади. Шунингдек, феъллар ҳам антонимик муносабатда бўлиши мумкин: *турмуш қурмоқ □ ажрашмоқ.*

RuWordNetда рус тилидаги от, сифат ва феъл сўз туркумидаги бирликлар киритилган. Масалан, **ДАВАТЬ (бермоқ)** феълига қуйидаги тушунчалар бириктирилган:

Sinset: *на руки выдать, дать, дать в руки, давать, отдать, выдать на руки, вручать, вручить, даваться, отдавать, передать, передавать*

Гипероним: *выдать, выдавать, преподнести, преподносить торжественно*

Гипоним: *вручать торжественно, вручить*

Гипоним: *подавать подавать в руки*

Гипоним: *недодать, недодавать*

Гипоним: *всунуть, всучить, всовывать, всучивать*

Гипоним: *купить, покупать, подкупать, подкупить*

Гипоним: *передать безвозмездно, передавать безвозмездно, безвозмездно передать, безвозмездно передавать*

Шунингдек, луғатда уядош (word similarity) ва семантик боғлиқ сўзлар (word relatedness) ҳам лингвистик база сифатида киритилади. Масалан, *ананас* ва *ўрик* ўзаро синоним сўзлар эмас, лекин улар мева туркумига кирганлиги учун битта парадигмага бирлашади ва уядош сўзларни ташкил қилади. Уларнинг таркибида эса ҳўл мева ва цитрус мева каби яна ички турлари мавжуд. Ўзаро семантик боғланган сўзлар эса бири иккинчисининг

келишини талаб қилади ҳамда ассоциатив хотирада улар ҳамиша ёнма-ён келади. Масалан, **пахта** сўзи~ *йиғмоқ,* ~ *терими*, ~ *дек юмшоқ* каби сўзлар билан валентлик ҳосил қилади. Бунда пахта юқоридаги лексемаларнинг ҳеч қайси бири билан маъно ўхшашлиги жиҳатидан боғлиқ эмас, лекин ёнма-ён келганда бир-бирини келишини талаб қилиб, муайян семантик гуруҳга бирлашади.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР

1. Абдураҳмонова Н. Компьютер лексикографиясининг айрим масалалари. — URL: https://conference.fledu.uz/
2. Christiane Fellbaum. Wordnet. An Electronic Lexical Database. 1998 Massachusetts Institute of Technology, P. 23–44
3. Atkins, B. T. S., and Levin, B. (1991). Admitting Impediments. In U. Zernik (Ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, 233–262.
4. Hillsdale, NJ: Erlbaum. Beckwith, R., Fellbaum, C, Gross, D., and Miller, G. A. (1991). WordNet: A Lexical Database Organized on Psycholinguistic Principles. In U. Zernik (Ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, 211–232.
5. Berlin, B., Breedlove, D., and Raven, P. H. (1973). General Principles of Classification and Nomenclature in Folk Biology. American Anthropologist, 75, 214– 42.
6. https://wordnet.princeton.edu

REFERENCES

1. Abduraҳmonova N., *Komp'yuter leksikografiyasining airim masalalari* (Some issues of computer lexicography) https://conference.fledu.uz/
2. Christiane Fellbaum., *Wordnet. An Electronic Lexical Database*. Massachusetts: Massachusetts Institute of Technology, 1998. 447 p.
3. Beryl Atkins and Beth Levin, *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon,* Hove and London: Lawrence Erlbaum Associates, 1991, 429 p.
4. Hillsdale, NJ: Erlbaum. Beckwith, R., Fellbaum, C, Gross, D., and Miller, G. A. (*WordNet: A Lexical Database Organized on Psycholinguistic Principles*. Lexical Acquisition: Exploiting On-line Resources to build a lexicon, 1991, 232 p.
5. Berlin Brent, Breedlove Dennis E., Raven Peter H., *General Principles of Classification and Nomenclature in Folk Biology,* American Anthropologist, 1973. 242 p. http://hdl.handle.net/10822/536563
6. https://wordnet.princeton.edu