Български език / Bulgarian Language 65 (2018), 2, 59–76

Представяне на знания чрез семантични мрежи

Светлозара Лесева
Институт за български език „Проф. Л. Андрейчин“ при БАН
zarka@dcl.bas.bg

Knowledge Representation by means of Semantic Nets

Svetlozara Leseva
IInstitute for Bulgarian Language *Prof. L. Andreychin*
Bulgarian Academy of Sciences
zarka@dcl.bas.bg

This paper presents an overview of existing semantic nets and major trends in their development which include integration of existing and emerging resources and databases in uniform systems whose backbone is formed by a taxonomy and/or an ontology in such a way as to achieve large coverage (in terms of concepts, instances, facts, etc.) coupled with the capability to perform complex inferences based on the axiomatisation of knowledge.

Keywords: semantic nets, ontologies, knowledge bases

1. Въведение

Семантичните мрежи са форма на представяне на човешкото познание във вид на насочен мрежовиден граф, в който възлите са концептуални обекти – понятия или конкретни същности (имена, дати и др.), а насочените дъги помежду им изразяват различни семантични релации (свойства), свързващи съответната двойка обекти (Lehmann 1992). Въз основа на горната дефиниция към семантичните мрежи може да се отнесат и ресурси, които не са замислени като такива, но са придобили подобни характеристики в резултат на разширяването и интегрирането им с други ресурси.

Едни от най-ярките примери за семантични мрежи са Уърднет (Word Net), ФреймНет (FrameNet), БейбълнНет (BabelNet), Уикипедия (Wikipedia), ДибиПедия (DbPedia), Фрийбейз (Freebase). Тенденцията е тези мрежи непрекъснато да се разширяват и интегрират помежду си и с други колекции от знания, което води до еволюция на самата концепция за създаваните ресурси.

Семантичните мрежи организират обектите от определена област на човешкото познание по начин, отразяващ отношенията помежду им, а оттам и позволяващ извършването на различни видове изводи за езиковите и/или извънезиковите обекти. Съществуват различни видове мрежи (Sowa

1992), от които ще посочим следните: 1) дефиниционни, или таксономични (definitional networks) – организират концептите преди всичко чрез релацията *род – вид*, като по-конкретните концепти транзитивно наследяват свойствата на по-абстрактните; 2) пропозиционални (assertional networks) – моделират съждения, т.е. отношения между предикати и аргументи; 3) импликационни (implicational networks) – организирани са чрез релации като логическо следване или каузалност. В реални приложения, ориентирани към представянето на човешкото знание и организацията му в семантичната памет, различните типове мрежи на практика се съчетават.

В настоящото изложение фокусът ще бъде поставен върху лингвистични ресурси, които се причисляват към семантичните мрежи, и върху тяхното надграждане в семантични мрежи от по-висок порядък чрез интеграцията им с енциклопедични, онтологични и други ресурси.

2. Лингвистичните ресурси като семантични мрежи

В тази част ще бъдат разгледани два класически лингвистични ресурса, в чиято концепция е залегнала идеята за мрежовидна организация – Уърднет (букв. „мрежа от думи“) и ФреймНет (букв. „мрежа от фреймове“). Тъй като целта им е да организират лексикалната система на даден език в единна релационна структура, те могат да се нарекат още и лексикално-семантични мрежи.

2.1. Уърднет като семантична мрежа

Лексикално-семантичната мрежа Уърднет[1] (Miller 1995, Fellbaum 1998) е най-широко използваният и развиван лексикален ресурс от този тип. Възлите в Уърднет представляват синонимни множества, а свързващите ги дъги са различни видове релации: 1. концептуални, отнасящи се до самите понятия и реализирани между синонимни множества, например отношението *род* (*хиперонимия*) – *вид* (*хипонимия*); лексикални – между двойки членове на синонимни множества (например *антонимия*); 2. деривационни, част от които са с експлицирана семантика, като *пиша – Агент – писател, рисувам – Събитие – рисуване*; 3. извънезикови и метаезикови релации, които изразяват тематична или регионална принадлежност или особености на употребата.

Изчерпателно представяне на структурата на Уърднет и на релациите и техните свойства в ракурса на Българския уърднет (БулНет) е представено у Коева (Коева/Koeva 2014).

Уърднет е показателен по отношение на съчетаването на свойствата на различни типове семантични мрежи. Голяма част от лексиката (съществителните имена) е структурирана в лексикална йерархия на наследяване (Miller 1990) чрез таксономичната релация *хиперонимия* и нейната обратна релация *хипонимия*. Друга таксономична релация е *холонимията* и обратната ѝ релация *меронимия*, изразяващи отношението *цяло – част*. Уърднет

споделя свойства и с импликационните мрежи чрез включването на езикови еквиваленти на импликацията (entailment), които служат за организация на системата на глаголите. Разглеждат се четири вида импликация (Fellbaum 1990): а) *строго включване* – протичането на ситуацията, описвана от имплициращия глагол, се включва времево в ситуацията, описвана от имплицирания глагол: *хъркам – спя;* б) *тропонимия (коекстензивност)* – имплициращият глагол е начин на извършване спрямо имплицирания и двете ситуации съвпадат времево: *шепна – говоря;* в) *обратна пресупозиция* – имплициращият глагол предполага предварителното извършване на действието, означено от имплицирания глагол: *развързвам – връзвам*; г) *каузация* – имплициращият глагол означава действие, което предизвиква резултат, означен от имплицирания глагол: *давам – имам*. Релациите в Уърднет, съответстващи на връзки в пропозиционалните семантични мрежи, кореспондират с тематични релации като *Агент*: *пиша – писател, Инструмент*: *мета – метла* и др. Макар интегрирането на кореспондиращи със семантични роли релации да е дискутирано още у Милър (Miller 1990), то не е проведено системно на семантично равнище. Включените в Уърднет релации от този тип се основават на деривационни отношения в английския език (Fellbaum 2009) и са реализирани чрез приписването на семантика на деривационната релация. С цел преодоляването на това ограничение са предприети инициативи за интегрирането на тематични отношения чрез съотнасянето на синонимните множества с единици от други лексикални ресурси, каквито са ФреймНет и ВърбНет (VerbNet), които отразяват именно този аспект на лексикалната семантика.

Релация, която сродява Уърднет с онтологиите, е *хипоним собствено име* (*instance hyponym*), чрез която се въвеждат синонимни множества, означаващи не понятия, а единични обекти, например: *Албер Камю – писател*. Включването на такива синонимни множества е аналог на „населването“ на онтологиите с екземпляри. Именно това е една от тенденциите за обогатяване на структурата на Уърднет особено в контекста на интеграцията с Уикипедия, в която много от статиите са именно екземпляри на класове.

С оглед на използването им за целите на автоматичната обработка на естествения език, семантичните мрежи като Уърднет най-често еволюират и се надстрояват чрез: а) разширяване и надграждане в рамките на съответния език; б) интегриране с уърднети за други езици. Вътреезиковото обогатяване в общия случай се осъществява чрез включване на нови концепти и единични обекти, дефиниране на нови релации и разширяване със съществуващи релации. Инициативи, насочени към разработването на съотносими ресурси за различни езици, са предприети в рамките на проекти като ЮроУърдНет (EuroWordNet, Vossen 1999), БалкаНет (BalkaNet, Tufis 2004) и др. Създаден е и така нареченият Глоубъл Уърднет Грид (Global WordNet Grid, Vossen 2016), чиято цел е да осигури еднозначно съотнася-

не на отделните уърднети и лексикална и компютърна оперативна съвместимост помежду им.

Силно изразена в последните години тенденция е и интегрирането на Уърднет с други лексикално-семантични, енциклопедични, онтологични и други ресурси (вж. т. 3).

2.2. ФреймНет като семантична мрежа

ФреймНет[2] (Ruppenhofer 2016) също е замислен като семантична мрежа, в която възлите са фреймове (рамки) или фреймови елементи. Фреймовете са описания на концептуални структури, представящи ситуации, обекти, събития от света, заедно с участниците в тях (аспекти или елементи на представяната ситуация, които се свързват с конкретни средства за езиково изразяване). Като част от даден фрейм участниците се наричат фреймови елементи (ФЕ) и се разделят на три вида: 1. *ядрени* – концептуално задължителни, необходими за интерпретирането на фрейма компоненти, които в специфичната си конфигурация правят фрейма уникален; 2. *периферни* – характеризиращи ситуацията от различни аспекти на протичането ѝ, без да въвеждат нови събития (включват свойства като време, място, начин, степен и т.н.); 3. *екстратематични* – концептуално принадлежат не към фрейма, в който са включени, а към друг абстрактен фрейм, като ситуират първия фрейм в по-широк контекст или осигуряват допълнителни детайли за ситуацията. Дадена лексикална единица „активира“ определен фрейм. Фреймовите елементи могат да се съотнесат косвено с тематичната структура, като най-общо аргументите съответстват на ядрени, а адюнктите – на периферни и/или екстратематични елементи. Съотнасянето между фреймовите елементи и позициите в синтактичната структура се заявява декларативно в анотираните примери за употреба на лексикалните единици, като за даден фреймов елемент се посочва граматичната му функция и фразовата му категория. Така например фреймът *Консумиране*[3] се формулира по следния начин: „Консумиращият поема храна или напитка *(Консумирано вещество*) през устата в храносмилателната си система. Възможно е да използва *Инструмент*...“. Ядрените елементи са *Консумиращ* и *Консумиран обект*. Периферни елементи са *Инструмент*, *Продължителност*, *Източник*, *Място*, *Време* и др. Възможна е и реализацията на екстратематични елементи като *Копартиципант*. Анотацията на примерите има следния вид:

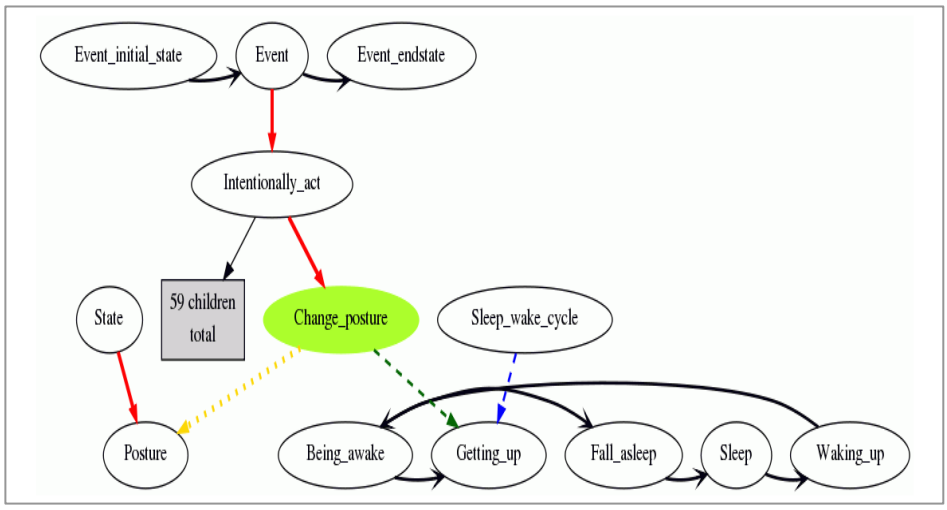
[Всеки ден]$_{\text{ВРЕМЕ}}$ [в Борисовата градина]$_{\text{МЯСТО}}$ [Иван]$_{\text{КОНСУМИРАЩ}}$ яде [сладолед]$_{\text{КОНСУМИРАН ОБЕКТ}}$ [от будката]$_{\text{ИЗТОЧНИК}}$ [с лъжичка]$_{\text{ИНСТРУМЕНТ}}$ [в продължение на часове]$_{\text{ПРОДЪЛЖИТЕЛНОСТ}}$ [заедно с Мария]$_{\text{КОПАРТИЦИПАНТ}}$.

Тъй като структурирането на познанието във ФреймНет отразява семантиката на ситуациите, то този тип ресурс споделя свойствата на пропозиционалните семантични мрежи.

ФреймНет е вътрешно организиран в мрежа от релации между концептуалните фреймове, част от които са онагледени на Фигура 1. чрез фрейма *Промяна_на_позата* (Change_posture): *Наследяване* – даден фрейм е подтип на надредния фрейм. Например *Промяна_на_позата* е наследник на *Действам_съзнателно* (Intentionally_act); *Използва* – подредният фрейм предполага или е поставен в контекста на надредния – *Ставане* (Getting_up) *Използва Промяна_на_позата*; *Подфрейм* – подредният фрейм е съставна част на по-сложно събитие, представено от надредната рамка, например *Ставане, Будно_състояние* (Being_awake), *Заспиване* (Fall_asleep), *Спане* (Sleep), *Събуждане* (Waking_up) са подфреймове на *Цикъл_сън_будуване* (Sleep_wake_cycle); *Предхожда* – даден фрейм предхожда във времето друг фрейм, заедно с който участва в по-комплексно събитие, например *Спане* (Sleep) *Предхожда Събуждане* (Waking_up) и *Се_предхожда_от Заспиване* (Fall_asleep).

Освен тях във ФреймНет са дефинирани още релациите: *Перспектива на* – подредният фрейм представя определена перспектива на по-абстрактен фрейм (например *Сценарий_заетост* (Employment_scenario) се перспек-

Фиг. 1. Графично представяне на подструктурата от релации за фрейма *Промяна на позата* в програмата за визуализация FrameGrapher[4] на проекта *ФреймНет*



тивизира от *Сценарий_работодател* (Employer_scenario) и *Сценарий_служител* (Employee_scenario); релациите *Каузативен_на* и *Инхоативен_на* обозначават разлики в събитийната структура (например *Предизвиквам_изменение_по_скала* (Cause_change_position_on_a_scale) е каузативен спрямо *Изменение_по_скала* (Change_position_on_a_scale), а *Изменение_*

по_скала е инхоативен спрямо *Позиция_върху_скала* (Position_on_ a_scale). Виж_също означава отношение на сходство между близки, но притежаващи значими различия фреймове (например *Поставяне* (Placing) и *Напълване* (Filling).

При все че релационната структура между фреймовете не е изцяло разгърната – част от тях не са свързани с други, налице са „празнини“ в структурата и под. – този тип информация увеличава значително използваемостта и ценността на ФреймНет, а обогатяването и усъвършенстването ѝ е една от посоките за развитие и интеграция с други ресурси.

3. Интеграция на семантични мрежи и други ресурси

В идеалния случай интеграцията между ресурси цели не просто обединяването на колекции от данни, така че да се получи по-голяма по обем колекция, а и придаването на допълнителна стойност на получения масив чрез съчетаване на преимуществата и неутрализирането на недостатъците на изходните ресурси и придобиването на качествено нови характеристики.

3.1. Интеграция между лексикално-семантични мрежи

По отношение един на друг Уърднет и ФреймНет представляват естествен избор на ресурс с цел взаимна интеграция. Обединяването им предлага възможност за съчетаването на богатата релационна структура, представяща лексикалната система чрез разнообразни релации между синонимни множества или лексеми, и детайлната концептуална структура на лексемите, съпътствана с информация за синтактичното ѝ изразяване.

По същността си интеграцията между Уърднет и ФреймНет се състои в съотнасяне на синонимни множества с кореспондиращи им фреймове. Така например предложеният от Тонели и Пигин (Tonelli 2009) подход извършва съотнасяне между двойки от вида <лексикална единица от ФреймНет : член на синонимно множество от Уърднет>, като се вземат предвид редица езикови фактори, включително съответствието между двойки (английски и италиански) лексикални единици, които принадлежат към еквивалентни синонимни множества в МултиУърднет (MultiWordNet[5]), и двойки (английски и италиански) лексикални единици, анотирани с един и същи фрейм в паралелния английско-италиански корпус, с приписани фреймове от ФреймНет. Други предлагани разработки за свързване между семантични мрежи включват съотнасянето между Уърднет, ФреймНет и ВърбНет[6] (Shi 2005); системата СемЛинк (SemLink[7]), в която се съотнасят Уърднет, ФреймНет, ВърбНет и Пропбанк (Propbank, Palmer 2009), както и СемЛинк + (SemLink +), включваща ФреймНет, ВърбНет, Пропбанк и Онтоноутс (Ontonotes[8]) (Palmer 2014).

3.2. Интеграция с Уикипедия

Макар да не е замислена като такава, многоезичната свободна енциклопедия Уикипедия[9] представлява семантична мрежа, в която възлите са статиите (под формата на интернет страници), а релациите са различните видове препратки: между статии с еднакво заглавие на различни езици (*междуезикови връзки*); между понятие, споменато в дадена статия, и съответстващата му статия в Уикипедия (*вътрешни препратки*); между статия и една или повече тематични категории (*тематични препратки*); между статии със синонимни или свързани заглавия (*пренасочващи препратки*).

Семантичната свързаност между понятията в Уикипедия се разбира в много по-широк план в сравнение с ресурси като Уърднет и обхваща богато множество от семантични релации. Сред тях са както отношения като *синонимия*, *хиперонимия*, *антонимия*, *меронимия* и т.н., така и такива като *употреба*, *функция*, *произход*, *способност* и много други. При все че семантичното описание в Уикипедия е богато и разностранно и с много поголяма гъстота на релациите в сравнение с Уърднет, то се характеризира с относително слаба структурираност, липса на ясна йерахична организация, липса на етикети на релациите.

Път за преодоляване на този проблем е интегрирането на лексикографските и енциклопедичните знания от Уикипедия с познанието, представено в Уърднет, в онтологии или под. В резултат от това възникват нов тип семантични мрежи като многоезиковата БейбълнНет (Navigli 2012), бази от знания като YAGO (Suchanek 2007), ДибиПедия (Lehman 2012) и други. По същността си съотнасянето между Уърднет и Уикипедия е насочено към преструктурирането на концептите и единичните обекти в Уикипедия в таксономия, като се използва най-вече хиперонимно-хипонимната релация в Уърднет. По-долу е разгледано интегрирането на Уикипедия и Уърднет в рамките на системата БейбълнНет.

3.2.1. БейбълнНет

БейбълнНет (BabelNet[10]) се самоопределя като съчетание от многоезичен енциклопедичен речник и семантична мрежа, създадена чрез взаимната интеграция на Уърднет и Уикипедия, както и на множество други лексикални ресурси[11], сред които Уикиречник (Wiktionary[12]), Уикиданни (Wikidata[13]), Отворен многоезиков уърднет (Open Multilingual WordNet[14]), Уикицитати (Wikiquote[15]), ВърбНет, ФреймНет, Терминологията на Майкрософт (Microsoft Terminology[16]), базата данни от географски имена Джионеймс (GeoNames[17]), базата данни с изображения Имиджнет (ImageNet[18]) и много други. Интеграцията е извършена чрез прилагането на алгоритъм за автоматично съотнасяне и чрез автоматичен превод на липсващите възли в даден език. Актуалната версия 4.0. съдържа данни за 284 езика, които са организирани в близо 16 милиона многоезикови синонимни множества (Бейбъл синсети – Babel synsets), свързани помежду си с над 1,3 милиарда лексикално-семантични релации[19].

Знанието в БейбълнНет е представено под формата на етикетиран насочен граф, в който всеки възел включва множество от лексикализации на даден концепт или единичен обект в различни езици например {play$_{en}$, Theaterstück$_{de}$, dramma$_{it}$, obra$_{es}$, pièce de théâtre$_{fr}$, пиеса$_{bg}$,...}, които съставят така наречения *Бейбъл синсет* (Navigli 2012). Концептите и релациите включват съществуващите в Принстънския уърднет и Уикипедия. По-конкретно – концептите от Уърднет са всички значения (синонимни множества) в него, а релациите са лексикалните и семантичните връзки (релации) между значенията; концептите от Уикипедия представляват самите енциклопедични страници, а хиперлинковете са преформулирани като неконкретизирани релации на свързаност (relatedness). Двата ресурса са обединени чрез автоматично съотнасяне между значенията в Уърднет и уики-страниците. Многоезиковите лексикализации се извличат от преводните еквиваленти на страниците чрез междуезиковите линкове, както и чрез автоматичен превод на реализациите на значенията в семантично анотирани корпуси.

3.3. Интеграция на семантични мрежи с онтологии

Интеграцията на лингвистични и други ресурси с онтологии цели надграждането на тези ресурси в две основни посоки: а) добавяне на абстрактно ниво на описание на концептуалното и/или фактологичното познание чрез съотнасянето с класовете на онтологията (вж. по-долу); б) аксиоматизация на съдържащото се познание чрез формулиране на специални правила (аксиоми), които ограничават интерпретацията на термините и правят възможно извеждането на по-сложни съждения.

3.3.1. Онтологии

Онтологиите са структури от знания за една или повече области от човешкото познание, представени формално чрез набор от: а) *термини* – класове (концепти), техни екземпляри, релации между концепти, функции и други обекти, които съставят речника, описващ съответната област; б) *дефиниции*, които съотнасят наименованията на термините с описание на значението им; в) *формални аксиоми*, които ограничават интерпретацията и правилната употреба на термините (Gruber 1995).

Според степента на специфичност онтологиите се делят на общи (foundational ontologies, upper-level ontologies, top-level ontologies, generic ontologies), междинни (core ontologies, middle-level ontologies) и предметни (domain ontologies) (Oberle 2006). В съвременния си вариант много общи онтологии са разширени с по-конкретни нива на концептуално описание и представляват модули в по-обхватни структури – онтологии или бази от знания, като при необходимост могат да се използват и самостоятелно.

Отчетлива тенденция представлява интегрирането на вече разширените онтологии с други ресурси, включително лексикално-семантични (на-

пример Уърднет), или с масиви от електронно съдържание, каквито са Уикипедия и различните бази от знания, извлечени от нея (като YAGO[20] и ДибиПедия), или създадени въз основа на други източници (Фрийбейз, Джионеймс и т.н.).

3.3.1.1. SUMO

В първоначалния си вид SUMO (Suggested Upper Merged Ontology)[21] (Niles 2001, Pease 2011) представлява формална обща онтология, базирана на предикатна логика от първи ред и създадена чрез обединяването на съществуващи свободно достъпни онтологии в единна и изчерпателна структура. Впоследствие тя е разширена с междинната онтология (MILO – Mid-Level Ontology) и допълнително свързана с множество предметни онтологии, колекции от данни и т.н.

Най-абстрактното ниво на общата онтология на SUMO е съставено от множество от абстрактни концепти. Началният възел *Същност (Entity)*[22] се разделя на класовете *Физическа същност (Physical)* и *Абстрактна същност (Abstract)*. Разграничението между *Процеси (Process)* и *Обекти (Object)* се извършва на следващото ниво на конкретизация като подкласове на *Физическа същност*. *Абстрактна същност* включва концепти като *Атрибут (Attribute)*, *Количество (Quantity), Клас (Class), Множество (Set), Граф (Graph)*. Като листа, наследници на *Обект,* в онтологията са включени концепти за животни и растения, съответстващи на биологични класове (*Птица, Влечуго*), семейства (*Котки*), видове (*Човек*) и т.н., а като наследници на *Процес* – конкретни действия и процеси като пиене *(Drink(ing)),* ядене *(Eat(ing))* и т.н.

SUMO включва: (i) оригиналната обща онтология, съдържаща около 1000 класа, 4000 аксиоми и 750 правила (Bond 2014); (ii) междинната онтология MILO – няколко хиляди допълнителни термина от по-ниско ниво и аксиоми, които ги дефинират; (iii) няколко десетки хиляди предметни онтологии от различни области на познанието. Заедно с разширенията си SUMO съдържа около 25 000 термина с богат набор от аксиоми (около 80 000), правила и релации, като всички термини са формално дефинирани.

Термините (по-конкретно концептите и екземплярите) в онтологията се асоциират с разнообразна информация:

1) Дефиниция – кратко описание на значението, в което може да се съдържат препратки към класове, релации и др. Например концептът *Human*[23] (*Човек*) е представен със следната дефиниция: *Modern man, the only remaining species of the Homo genus* (*съвременен човек, единственият съществуващ днес вид на рода Homo*).

2) Илюстративен материал, онагледяващ термина, който включва препратки към изображения от Уикипедия или от други източници.

3) Дизюнктивни термини – класове, с които терминът няма общи екземпляри. Примери за дизюнктивни термини за *Human* са термините *Or-*

ganization (Организация) и *DomesticAnimal (Домашно животно)*. Формулирането на отношението на дизюнкция е чрез предикат *дизюнктивни (disjoint)*, на който термините са аргументи.

(4) Надкласове на термина – класове, които включват термина, например: *CognitiveAgent* (*Познавателен агент*) и *Hominid* (*Хоминид*) по отношение на *Human.* Формулирането на отношението между клас и негов надклас се извършва чрез предиката *подклас* (*subclass*), на който терминът и неговият надклас са съответно първи и втори аргумент.

(5) Подкласове на термина – класове, които се включват в класа, дефиниран от термина, например: *HumanAdult* (*Възрастен човек*)*, Woman* (*Жена*)*, Man* (*Мъж*)*, Teenager* (*Юноша*) по отношение на *Human.* Формулирането на отношението между клас и негов подклас се извършва чрез предиката *подклас* (*subclass*), на който терминът и неговият подклас са съответно втори и първи аргумент.

(6) Други релации, в които терминът е аргумент – релации, дефинирани в онтологията (включително и в MILO и тематичните онтологии), в които даденият концепт участва като аргумент, например: *гражданин* (*човек, нация*).

(7) Класове от MILO и предметните онтологии, които са еквивалентни на съответния концепт в SUMO или представляват негови подкласове, надкласове, екземпляри и т.н.

(8) Аксиоми – включват ограничения, валидни за разглеждания термин, извлечени както от общата онтология, така и от MILO и от интегрираните предметни онтологии, например:

Ако **човек** е екземпляр на класа *Човек* и
организация е екземпляр на класа *Организация* и
позиция е екземпляр на класа *Позиция* и
ролята **член** на **организация** и ролята **позиция** са свойства на **човека**,
тогава екземплярът **човек** заема **позиция** в **организацията**

SUMO е съотнесена с Уърднет, YAGO и други ресурси, което, доколкото ни е известно, я прави единствената онтология, интегрирана пълно и последователно с лексикално-семантичната мрежа и с база от знания.

3.3.2. Съотнасяне на SUMO с Уърднет

Съотнасянето между SUMO и Уърднет представлява асоцииране на всяко синонимно множество в Уърднет с концепт в разширената SUMO. Възможните начини за съотнасяне са три (De Melo 2008):

1) Чрез релация на еквивалентност между синонимно множество и даден концепт в онтологията, например множеството {discipline, subject, subject area, subject field, field, field of study, study, bailiwick} е еквивалентно с концепта *Дисциплина (Field of study)*[24].

2) Чрез релация на включване между по-общ концепт в онтологията и по-конкретно лексикално значение (хипоним) в Уърднет – например синонимното множество {differential calculus} (диференциално смятане)[25] се съотнася с по-общия концепт *Дисциплина* в SUMO.

3) Чрез релация на инстанциация (между *екземпляр* в Уърднет и концепт в SUMO). По силата на такава релация синонимното множество {George Washington, President Washington Washington} се съотнася с концепта, на който е екземпляр – класа *Man* (*Мъж*)[26].

Чрез интегрирането с Уърднет онтологията е допълнително обогатена с голям брой таксономично организирани подкласове на съществуващи класове и техни екземпляри, с езикови релации, които не са представени в онтологията. Тъй като Уърднет е разработен за множество езици, онтологията може да се използва като многоезиков ресурс.

3.3.3. YAGO

YAGO представлява голяма семантична база от знания, която обединява енциклопедичната и езиковата информация от Уикипедия, Уърднет и Джионеймс и включва над 10 милиона концепта и техни екземпляри и над 120 милиона факта (около 450 милиона заедно Джионеймс), които ги характеризират[27].

3.3.3.1. Организация на YAGO

YAGO е конструирана автоматично, като същностите и фактите са извлечени от системата от категории и информационните шаблони (infobox) в Уикипедия (Hoffart 2013). Фактите представляват тройки от вида < *същност : релация : същност* >, като за всяка статия от Уикипедия в базата от знания се създава отделна същност. Екземплярите на класовете се съотнасят с класове в Уикипедия, след което се търси съответствието на класовете от Уикипедия в Уърднет. Например в Уикипедия статията *ИванВазов* (IvanVazov) е класифицирана в няколко категории, сред които *Български поети, Български романисти, Български писатели, Български драматурзи.* На базата на тази информация в YAGO се създава същността *ИванВазов* и съответните класове, на който тя е екземпляр – *Български поет*, *Български романист, Български писател, Български драматург*. На следващата стъпка класът *Български поет* се определя като подклас на класа *Поет*, за който е установено съответствие със синонимното множество {poet:1} в Уърднет. По аналогичен начин се съотнасят и останалите класове. Връзката с Уърднет превръща базата от знания в таксономия, в която концептуалната и лексикалната информация са организирани йерархично чрез хиперонимно-хипонимната релация. В резултат от съотнасянето с Уърднет концептите са обединени в повече от 350 000 класа.

В YAGO ръчно са дефинирани около 100 релации, които характеризират включените в базата от знания същности, например: *е_носител_ на_*

награда, е_роден_в, е_роден_на_дата, има_столица. За извличане на екземпляри на релациите са използвани категориите в Уикипедия и инфобоксовете на статиите, като са формулирани шаблони, които съотнасят категории и атрибути на инфобоксове с шаблони за факти. Шаблоните обхващат 200-те най-често срещани атрибута на инфобокса в Уикипедия. Така например атрибутът *роден=9 юли 1850* в статията ИванВазов се „превежда“ като *рождена дата ИванВазов (Ден 9 (Месец 7 (Година 1850)))*.

Версията YAGO2 включва и многоезикова информация, извлечена с помощта на междуезиковите връзки в Уикипедия и чрез преводните еквиваленти на синонимни множества на различни езици в Универсалния уърднет и Глоубъл Уърднет Грид. Най-актуалното развитие на базата от знания е свързано с мигриране към YAGO3 (Mahdisoltani 2015), където се включва информация от разноезични версии на Уикипедия.

3.3.3.2. Интегриране на YAGO2 със SUMO

Таксономичната организация на базата от знания, заимствана от Уърднет, е валидирана, интегрирана и обогатена спрямо SUMO, като се използва вече имплементираното съотнасяне между класовете в онтологията и синонимните множества. Точният пренос на таксономията на базата от знания към концептите на онтологията, а където се налага и дефинирането на нови класове, свързани непротиворечиво в йерархията на SUMO, се гарантира чрез алгоритъм, който търси съответствието на даден клас от YAGO в онтологията (De Melo 2008).

Поради специфичността на повечето екземпляри и класове, извлечени от Уикипедия, често в SUMO няма концепт, който да им съответства. В такъв случай съответният клас, например *Личности от Ливърпул*, се създава и се свързва като подклас на подходящ по-общ клас, извлечен от Уърднет, в случая класа *Лице*. След това се проверява дали класът от Уърднет има корелат в SUMO и ако такъв корелат съществува, съотнасянето се извършва, например *Лице* (Уърднет) – *Човек* (SUMO). Ако синонимното множество няма еквивалент (например *небостъргач*), в SUMO се създава съответният нов клас *Небостъргач*, който се свързва със съществуващ надреден клас (в случая *Сграда*). По този начин еквивалентните и уникалните концепти от двата ресурса се обединяват в обща структура, в която се включват според нивото си на специфичност.

3.4. Бази от знания

Базите от знания представляват автоматично създадени от други източници ресурси, които съдържат огромно количество същности и факти, най-често организирани поне частично с помощта на специално създадена или съществуваща онтология или таксономия.

3.4.1. ДибиПедия

ДибиПедия е създадена чрез извличане на структурирана информация на базата на инфобоксовете и категориите в Уикипедия и включени в статиите географски координати, изображения и връзки към външни уеб страници, като по същество представлява RDF версия на съдържанието в Уикипедия. Системата позволява сложни заявки и свързване на различни колекции от данни в интернет с данни от Уикипедия.

Актуалната английска версия на ДибиПедия съдържа 6,6 милиона същности, на голяма част от които е приписана принадлежност към даден клас в плитка онтология, разработена за целите на инициативата[28], която се базира на най-често използваните информационни шаблони в Уикипедия. Тя обхваща множество тематични области и включва 685 класа[29] от категориите *Лица*, *Места, Творби, Организации, Биологични видове, Заболявания* и др., организирани йерархично чрез релацията на включване. Онтологията има вид на насочен ацикличен граф, като класовете може да имат повече от един надклас. Структурата може да се преобразува в таксономия, като се даде приоритет на първия суперклас, посочен в списъка. Онтологията е създадена чрез съотнасяне на елементи на инфобоксовете в Уикипедия с елементи (класове, свойства) на онтологията. По аналогичен начин е предприето и съотнасяне на версиите за различни от английския езици към споделената онтология, като се използват глобалните идентификатори от онтологията.

ДибиПедия предлага локализирани варианти за над сто езика (включително български). Актуалната версия от октомври 2016 година[30] съдържа 13 милиарда факта (RDF тройки), 1,7 милиарда от които са извлечени от английското издание на Уикипедия, 6,6 милиарда – от другоезични версии на Уикипедия, а 4,8 милиарда – от Уикипедия Комънс (Wikipedia Commons[31]) и Уикидейта. След добавянето на допълнителни данни за отделните езици общият брой на фактите достига 23 милиарда.

3.4.2. Съотнасяне с YAGO2

Съотнасянето между YAGO2 и ДибиПедия се извършва чрез няколко типа съответствия[32]:

(i) релации на идентичност между индивиди в двете бази от знания; (ii) релации на идентичност между класовете в YAGO2 и моделираните по модела на YAGO класове в ДибиПедия; (iii) релация *подклас_на* между класовете в YAGO2 и тези в онтологията на ДибиПедия; (iv) релация *подсвойство_на* между релациите в YAGO2 и свойствата в онтологията на ДибиПедия.

ДибиПедия включва връзки и към редица външни ресурси, включително към Уърднет, онтологии, бази от знания (Фрийбейз), бази от данни на музеи, статистически служби, проекта *Гутенберг* и множество други (Lehmann 2012).

4. Обобщение

Както показва представеният обзор, със създаването на нови лексикално-семантични, енциклопедични, онтологични и под. ресурси, в много случаи обединяващи и максимизиращи преимуществата на вече съществуващи ресурси, възникват нови типове колекции от данни, които са ориентирани към създаването на комплексен абстрактен модел на човешкото познание. Доминираща тенденция е тези обединени ресурси да съчетават концептуално моделираното описание, характерно за онтологиите, с огромния обем от факти в базите от знания. Като лексикално-семантична мрежа, покриваща голяма част от лексикалната система на даден език и представяща я в структуриран вид, Уърднет има средищно място – от една страна осигурява връзка и попълва „празнините“ в различните нива на онтологично представяне, а от друга – подобрява лексикално покритие на описваните концепти.

Независимо кои са изходните ресурси, данните в получените колекции обикновено се представят под формата на семантична мрежа. Това е естествен избор на структуриране на човешкото познание както защото семантичните мрежи наподобяват организацията на семантичната памет, така и защото представляват удобно представяне на информация с оглед на компютърната ѝ обработка.

Целта, методологията и крайният резултат може да бъдат различни, но общото е, че чрез обединяването на структурирани ресурси се постига: лексикално обогатяване и разширяване; взаимно допълване и верифициране на информацията; езиково обогатяване (при включването на повече от един език); таксономизация на лексикалната информация; създаване на по-гъста мрежа от релации.

Една от основните посоки за изследване в областта на семантичните мрежи е откриването на начини за установяване на семантиката на асоциативните връзки в Уикипедия, включването им в общата система на семантичните отношения, формалното им дефиниране и кодирането им във взаимосвързаните ресурси. Друго широко изследователско поле представлява интегрирането на аксиомите от онтологиите, тъй като това увеличава значително възможността за автоматичното извършване на по-сложни съждения на базата на наличната информация.

Благодарности

Изследването е извършено в рамките на проекта „Семантична мрежа с широк спектър от семантични релации“, подкрепен от Фонд „Научни изследвания“ по програма „Финансиране на фундаментални научни изследвания“, Договор No. 10/39/2016.

БЕЛЕЖКИ / NOTES

[1] https://wordnet.princeton.edu/

[2] https://framenet.icsi.berkeley.edu/fndrupal/

[3] https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Ingestion

[4] Графично представяне на фрейма Change_posture и релациите, в които участва: https://framenet.icsi.berkeley.edu/fndrupal/FrameGrapher

[5] http://multiwordnet.fbk.eu/english/home.php

[6] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[7] https://verbs.colorado.edu/semlink/

[8] https://catalog.ldc.upenn.edu/ldc2013t19

[9] https://www.wikipedia.org/

[10] http://babelnet.org/

[11] http://babelnet.org/about

[12] https://www.wiktionary.org/

[13] https://www.wikidata.org/wiki/Wikidata:Main_Page

[14] http://compling.hss.ntu.edu.sg/omw/

[15] https://en.wikiquote.org/wiki/Main_Page

[16] https://www.microsoft.com/Language/en-US/Default.aspx

[17] http://www.geonames.org/

[18] http://www.image-net.org/

[19] http://babelnet.org/stats

[20] http://www.mpi-inf.mpg.de/yago-naga/yago/

[21] http://www. ontologyportal.org/

[22] http://www.ontology4.us/download/dot/SumoOntology.pdf

[23] http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?word=Human&POS=1

[24] http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?word=field+of+study&POS=1

[25] http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?word=differential+calculus&POS=1

[26] http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?word=George+Washington&POS=1

[27] Информацията е публикувана на адрес: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/, <30.06.2017 г.>

[28] http://wiki.dbpedia.org/services-resources/ontology

[29] http://mappings.dbpedia.org/server/ontology/classes/

[30] http://wiki.dbpedia.org/blog/new-dbpedia-release-%E2%80%93-2016-10

[31] https://commons.wikimedia.org/wiki/Main_Page

[32] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/linking/

ЛИТЕРАТУРА

Коева 2014: *Коева, Св.* Българският национален корпус в контекста на световната теория и практика. – В: Езикови ресурси и технологии за български език. София, Академично издателство „Проф. Марин Дринов“, 2014, с. 154–173.

REFERENCES

Bond 2014: *Bond, F., C. Fellbaum, S-K. Hsieh, C-R Huang, A. Pease, P. Vossen.* A Multilingual Lexico-Semantic Database and Ontology. – In: Buitelaar P., P. Cimiano (eds.). Towards the Multilingual Semantic Web. Springer-Verlag Berlin Hiedelberg, 2014, 243–258.

De Melo 2008: *De Melo, G., F. Suchanek, A. Pease.* Integrating YAGO into the Suggested Upper Merged Ontology. Technical report, MPI{I{2008{5{003.

Fellbaum 1990: *Fellbaum, C.* English Verbs as a Semantic Net. – *International* Journal of Lexicography, 1990 3 (4), 278–301.

Fellbaum 1998: *Fellbaum, C*. (ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

Fellbaum 2009: *Fellbaum, C., A. Osherson, P. E. Clark*. Putting Semantics into WordNet's "Morphosemantic" Links. – In: Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in HumanLanguage Technologies. Springer Lecture Notes in Informatics], vol. 5603, 350–358.

Gruber 1995: *Gruber, T. R.* Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies, Vol. 43, Issues 4–5, November 1995, 907–928.

Henrich 2012: *Henrich, V., E. Hinrichs, K. Suttner*. Automatically Linking Germa Net to Wikipedia for Harvesting Corpus Examples for GermaNet Senses. – Journal for Language Technology and Computational Linguistics (JLCL*)*, Vol. 27, No. 1, August 2012, 1–19.

Hoffart 2013: *Hoffart J., F. M. Suchanek, K. Berberich, G. Weikum*. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. – Artificial Intelligence (194), 28–61.

Koeva 2014: *Koeva, Sv.* Balgarskiyat natsionalen korpus v konteksta na svetovnata teoriya i praktika [The Bulgarian National Corps in the Context of World Theory and Practice]. – In: *Ezikovi resursi i tehnologii za balgarski ezik*. Sofia. Akademichno izdatelstvo Prof. Marin Drinov, 2014, 154–173.]

Lehmann 1992: *Lehmann, F.* Semantic Networks. – Computers and Mathematics with Applications, vol. 23, issue 2–5, January–March 1992, 1–50.

Lehmann 2012: *Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, Ch. Bizer.* DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. – Semantic Web Journal, 1, 2012, 1–5.

Mahdisoltani 2015: *Mahdisoltani, F., J. A. Biega, F. M. Suchanek.* Yago3: A Knowledge Base from Multilingual Wikipedias. – In: Conference on Innovative Data Systems Research (CIDR), 2015 (ръкопис).

Miller 1990: *Miller, G. A*. Nouns in WordNet: A Lexical Inheritance System. – International Journal of Lexicography, (1990), 3 (4), 245–264.

Miller 1995: *Miller, G. A*. WordNet: A Lexical Database for English. – Communications of the ACM, Vol. 38, 1995, No. 11, 39–41.

Navigli 2012: *Navigli, R., S. P. Ponzetto*. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. – Journal of Artificial Intelligence, vol. 193, Elsevier Science Publishing, 2012.

Niles 2001: *Niles, I., A. Pease.* Towards a Standard Upper Ontology. – In: Welty, C., B. Smith (eds.). Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). ACM Press, 2001, 2–9.

Oberle 2006: *Oberle, D.* Semantic Management of Middleware. Springer, 2006.

Palmer 2009: *Palmer, M*. Semlink: Linking Propbank, VerbNet and FrameNet. – In: Proceedings of the Generative Lexicon Conference, 9–15.

Palmer 2014: *Palmer, M., C. Bonial, D. McCarthy*. SemLink+: FrameNet, VerbNet and Event Ontologies. – In: Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), Baltimore, Maryland USA, June 27, 2014. ACL 2014, 13–17.

Pease 2011: *Pease, A*. Ontology: A Practical Guide. Articulate Software Press, Angwin, CA, 2011.

Ponzetto 2009: *Ponzetto, S. P., Navigli, R*. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. – In: AAAI Publications, Twenty-First International Joint Conference on Artificial Intelligence, 2083–2088.

Ruppenhofer 2016: *Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, Ch. R., Baker C. F., Scheffczyk, J*. FrameNet II: Extended Theory and Practice (Revised November 1, 2016).

Shi 2005: *Shi, L., R. Mihalcea*. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. – In: Cicling, Mexico, 2005.

Suchanek et al. 2007: *Suchanek, F. M., G. Kasneci, G. Weikum*. Yago: A Core of Semantic Knowledge. – In: Proceedings of the 16th international conference on World Wide Web, WWW '07, 697–706, New York, NY, USA. ACM.

Tonelli 2009: *Tonelli, S., D. Pighin*. New Features for FrameNet – Wordnet Mapping. – In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09), Boulder, CO, USA, 2009.

Tufiș 2004: *Tufiș, D., D. Cristea, S. Stamou*. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. – Romanian Journal of Information Science and Technology, Special Issue, 7, 1–2, 9–43.

Vossen 1999: *Vossen, P*. (ed.). EuroWordNet: a Multilingual Database with Lexical Semantic Networks for European Languages. Kluwer Academic Publishers, Dordrecht.

Vossen 2016: *Vossen, P., F. Bond, J. McCrae*. Toward a Truly Multilingual Global Wordnet Grid. – In: Proceedings of the Eighth Global Wordnet Conference, Bucharest, Romania, 27–30 January. Bucharest: Research Institute for Artificial Intelligence – Romanian Academy, 419–426.

РЕЗЮМЕ

В статията се представя обзор на съществуващите семантични мрежи и основните тенденции в тяхното развитие, които най-общо се изразяват в интегриране на

съществуващи и нововъзникващи ресурси и бази от данни в единни системи, структурирани въз основа на таксономии и/или онтологии. Стремежът е наред с голямото покритие на данните да се постигне аксиоматизация на познанието, което да позволява извършване на сложни съждения.

Ключови думи: семантични мрежи, онтологии, бази от знания

✉ *Гл. ас. д-р Светлозара Лесева*
Секция по компютърна лингвистика
Институт за български език „Проф. Л. Андрейчин“ при БАН
бул. „Шипченски проход“ 52, бл. 17, 1113 София, България

✉*Assist. Prof. Svetlozara Leseva, PhD*
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, 1113 Sofia, Bulgaria