ТЕРМИНОЛОГИЈА И ДИГИТАЛИЗАЦИЈА

Интервју са проф. др Ранком Станковић

Ранка Станковић је ванредни професор на Универзитету у Београду, на Катедри за примењену математику и информатику Рударско-геолошког факултета где предаје више предмета из области информатике, статистике и геоинформатике. Такође предаје програмирање за лингвисте, екстракцију информација из текста и семантичког веба на мастер и докторским студијама Универзитета у Београду. Потпредседница је JERTEH-а, Српског друштва за језичке ресурсе и технологију, шеф рачунарског центра Рударско-геолошког факултета Универзитета у Београду и вођа тима за неколико националних софтверских пројеката. Низ година је председница Комисије за стандарде KS A037, Терминологија и својом енергијом и залагањем дала је посебан допринос активностима комисије које су резултовале објављивањем значајног броја стандарда. Међутим, оно што је посебно важно јесу активности на међународном нивоу, у оквиру рада техничког комитета ISO TC 37.

ИСС: Недавно сте објавили заједнички рад (Тања Ивановић, Ранка Станковић, Бранислава Шандрих Тодоровић, Цветана Крстева) о ресурсима и алатима који се користе за издвајање и евалуацију енглеско-српске терминологије у области енергетике. До каквих сте резултата дошли?

Р.С: У раду *Corpus-based bilingual terminology extraction in the power engineering domain* које је објавио угледни часопис *International Journal of Theoretical and Applied Issues in Specialized Communication: Terminology* представили смо систем *BiLTE* који смо развили за екстракцију двојезичних терминолошких парова из паралелних енглеско-српских корпуса и применили на корпус из домена електроенергетике. Систем користи језичке ресурсе и алате које већ деценијама развијају проф. др Душко Витас и проф. др Цветана Крстев са сарадницима окупљеним у Друштву за језичке ресурсе и технологије *JeRTeh*. Конкретно, *BiLTE* користи парални корпус у ком су поравнати преводи на нивоу реченица, српске морфолошке речнике и обрасце за екстракцију термина из српског дела корпуса са прецизношћу 86%, док се за енглески језик комбинују различити алати тако да је прецизност резултата спајања екстрактора за енглески 92%. Следећи корак је поравнање издвојених једнојезичних термина, где систем користи поравнате делове паралелног корпуса. Евалуација је показала да је прецизност издвајања двојезичног пара била од 70% до 72%, у зависности од примењене функције подударања. Осим развијеног система, конкретан резултат представљеног истраживања су 2 684 исправна двојезична пара која су ушла у терминолошку базу података Termi. Треба поменути да је осим у области електроенергетике, систем успешно коришћен и у другим областима, поменимо библиотекарство и геологију. Резултати показују да је систем могуће применити на било коју област, а предуслов је да постоје доступни преводи текстова из циљне области, који се потом паралелизују, након чега следи екстракција и поравнање термина.

ИСС: У раду се као један од извора терминологије из области енергетике помиње и Међународни електротехнички онлајн-речник Електропедија. Међу само двадесетак језика који имају своју терминологију из области електротехнике уопште од 2013. године Институт и експерти из комисија за стандарде увели су и српски језик са терминологијом из 45 области. Како бисмо могли да унапредимо рад на изради терминологије из преосталих области?

Р.С: Иако је Електропедија терминолошка база која је намењена да буде од помоћи у припреми стандарда, као и у њиховој примени, она пружа велику подршку професионалним преводиоцима, корисницима техничке литературе, у настави, писању техничких спецификација, као и у трговини. Електропедија је најсвеобухватнија онлајн-база термина у области електротехнологије која се континуално ревидира и проширује. Имајући у виду брзину којом настају нови термини који долазе са све бржим технолошким, нарочито ИТ развојем, свакако су развој репрезентативног корпуса и примена алата за екстракцију термина пут ка бржем развоју терминологије. Чињеница да Институт располаже великом количином двојезичних текстова је одлична основа да се крене у овакав важан подухват. Рад на развоју терминолошких база не треба схватити као једнократан посао, већ као континуиран или бар периодичан посао. Друштво *JeRTeh* располаже знањем, ресурсима и алатима којима може да подржи овај процес. Екстраховани парови термина су поређени са садржајем верзије Електропедије из 2018. године и пронађено је поклапање од 136 преводних еквивалената, што потврђује ваљаност нашег приступа, али показује и да постоји простор за допуну Електропедије резултатом наших истраживања. Савремени трендови развоја терминологије базирају се на корпусима текстова као извору терминологије и потврди употребе у контексту.

ИСС: Вештачка интелигенција је последњих година изузетно важна област у стандардизацији. У фокусу ваших интересовања је и вештачка интелигенција која је у вези са језичким технологијама. Каква је тренутна ситуација у тој области?

Р.С: Иако се изучава већ деценијама, вештачка интелигенција последњих година интензивно трансформише пословање данас. Аутономни уређаји и роботи налазе примену у различитим сценаријима и обучени су да остваре интеракцију и препознају окружење око себе. Модели дубоког учења обрађују огромне количине података, као што су фотографије, текстови и звуци, користећи вештачке неуронске мреже како би обезбедили тачне резултате.

Ја бих се ограничила у коментару на обраду природног језика као област вештачке интелигенције и лингвистике која се, између осталог, бави истраживањем аутоматског генерисања и разумевања природних људских језика. Једноставно говорећи, проблем је двосмеран: системи за генерисање природног језика претварају податке из формалних репрезентација у информације исказане природним језиком, док системи за разумевање природног језика претварају исказе на природном језику у формално структуриране податке које рачунарски програми могу да обраде.

Наш тим већ дужи низ година ради на развоју језичких модела различите намене: обележавање врста речи, препознавање такозваних именованих ентитета: особа, организација, локација, датума..., њихово повезивање са базама знања, однедавно и генерисање текста, а свакако важан сегмент јесте припрема ресурса у виду корпуса и лексичких база без којих нема развоја језичких модела.

ИСС: Између осталог, и потпредседница сте *JeRTeh*-а. Објасните нам које су основне активности друштва?

Р.С: Друштво *JeRTeh* је основано да би промовисало, популарисало и унапредило све гране језичких технологија на научном, стручном и практичном нивоу. Бавимо се осмишљавањем и реализацијом програма и пројеката везаних за развој језичких ресурса и алата, самостално или у сарадњи са другим институцијама.

Већ дуги низ година се редовно организује Семинар JeRTeh-а на ком су до сада излагали еминентни страни и домаћи истраживачи. Организујемо такође и радионице, саветовања и друге врсте јавних догађаја и разне облике едукације.

Као важну активност друштва бих поменула развој корпуса 100 старих српских романа који су први пут објављени у периоду од 1840. до 1920. године, насталих у оквиру COST акције CA16204 *Distant Reading for European Literary History* (удаљено читање за европску историју књижевности, од 2017. до 2022. године). Један од најважнијих циљева ове акције био је припрема вишејезичног корпуса (названог *European Literary Text Collection* – ELTeC) који је јавно доступан и слободан за употребу, снабдевен различитим лингвистичким информацијама, повезан са википодацима и пригодан за бројна даља лингвистичка, историјска и хуманистичка истраживања.

Отворени смо за сарадњу са образовним и истраживачким установама, удружењима, јавним органима, привредним друштвима и другим субјектима у земљи и иностранству који се баве сличним истраживањима, или који желе у пракси да примене неке од наших ресурса и алата.

ИСС: Дуго сте већ председница комисије за стандарде А037, *Терминолоīија* и последње две године били сте активни на међународном нивоу као вођа радне групе за измену једног изузетно важног стандарда за нас – ISO 12199. Ваше ангажовање допринело је да после више од двадесет година постигнемо да се објави ново издање поменутог стандарда. Реците нам нешто више о томе.

Р.С: Институт и друге институције у нашој земљи покушавале су више од двадесет година да реше проблем коришћења српског језика у појединим међународним стадардима. Конкретно, у питању је међународни стандард ISO 12199, *Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet,* усвојен у Међународном техничком комитету ISO/TC 37, Језик и терминологија, у ком је више пута покренуто питање измене назива српско-хрватски у српски језик, као и укључивање српског језика међу језике који користе латиничко писмо, које се поред ћирилице користи у српском језику.

Треба напоменути да поред напора и ангажовања наших представника на бројним онлајн-састанцима, дописа упућених различитим телима, успех не би био могућ да није било подршке и помоћи стручних служби Института и подршке Министарства спољних послова Републике Србије. Коначно је у јуну 2022. године објављено ново, односно ревидовано издање стандарда у коме је коначно спроведена захтевана измена https://www.iso.org/standard/84159.html.

ИСС: **Овогодишња сте добитница признања Института за допринос развоју националне стандардизације. Како бисте оценили значај рада у комисији за стандарде и шта бисмо могли да учинимо да се тај рад унапреди?**

Р.С: Када су језички ресурси у дигиталном облику, па тако и терминологија у питању, нажалост морамо констатовати да Србија значајно заостаје не само за такозваним „великим" језицима (енглески, француски, немачки, шпански) већ и за језицима из региона (словеначки, хрватски, бугарски). Мислим да је један од путева да се превазиђе, или бар смањи заостајање повезивање терминолога, лингвиста, лексикографа са истраживачима из области рачунарске лингвистике, и коначно и можда најбитније, индустрије и институције које би биле крајњи корисници и заинтересоване стране за технолошки напредна решења. Препознавање терминологије за специфичан домен је од кључног значаја за разумевање и превођење, било аутоматско или традиционално, али његова важност се такође огледа и у побољшању квалитета претраге код база података, база знања, различитих типова репозиторијума.

Важно је да развијени ресурси следе стандарде, како би се постигла могућност вишеструког коришћења и ефикасне размене података, потом информација, па можда и знања аутоматски. Управо ту видим улогу комисије која би могла да подигне свест о ISO стандардима и која поједина решења треба да следе. Радне групе које постоје у оквиру ISO/TC: 1) Принципи и методе, 2) Ток развоја терминологије и кодирање језика, 3) Управљање терминолошким ресурсима, 4) Управљање језичким ресурсима 5) Превођење, интерпретирање и повезане технологије, могу да понуде конкретне моделе и стандарде. Поменимо неколико важних примера стандарда: TMX (Translation memory exchange) за паралелне корпусе, TBX (TermBase eXchange) за моделе терминолошких база, LMF (Lexical Markup Framework) за моделе лексичких ресурса.

Припремила: Виолета Нешковић-Поповић