

Надугтенко Максим Вікторович

Український мовно-інформаційний фонд НАН України

ЗАГАЛЬНИЙ ОГЛЯД ТА ПЕРСПЕКТИВИ ВИКОРИСТАННЯ НАЦІОНАЛЬНИХ ЛІНГВІСТИЧНИХ ЦИФРОВИХ РЕСУРСІВ УКРАЇНСЬКОГО МОВНО-ІНФОРМАЦІЙНОГО ФОНДУ НАН УКРАЇНИ

Представлено короткий огляд існуючих національних лінгвістичних цифрових ресурсів Українського мовно-інформаційного фонду НАН України та окреслено основні перспективні напрямки застосування наробок Українського мовно-інформаційного фонду НАН України.

***Ключові слова:** віртуалізація, лексикографічна система, електронний словник, хмаринне середовище, лінгвістичні ресурси, лінгвістична інфраструктура*

Основні напрямки роботи Українського мовно-інформаційного фонду НАН України: дослідження системної будови природної мови; теоретична, математична і прикладна лінгвістика; створення та ведення фундаментального корпусу української мови та інших мов; розробка і створення інформаційно-лінгвістичних систем (традиційних і комп'ютерних словників, підручників, навчальних, експертних систем, систем автоматизованого опрацювання природної мови та мережевих систем професійної взаємодії у лінгвістиці); дослідження лінгвістичних аспектів інтелектуальної діяльності; розробка, впровадження та експлуатація інформаційних систем, баз даних і знань загальнокультурного характеру.

Фахівцями Українського мовно-інформаційного фонду НАН України створено перший в нашій країні **Український національний лінгвістичний корпус**, який на сьогодні налічує близько 200 млн. слововживань і становить сучасну експериментальну основу для проведення фундаментальних мовознавчих студій та укладання лексикографічних праць нового покоління.

Український мовно-інформаційний фонд НАН України у 1994 році започаткував програму зі створення серії українськомовних словників нового

покоління – **«Словники України»**. «Словники України» є частиною єдиного у нашій державі наукового об'єкта, який має статус національного надбання – **Національної словникової бази України**, яка зараз налічує 92 друковані праці та 37 електронних систем.

Видатним досягненням Українського мовно-інформаційного фонду є створення та видання одинадцяти версій першого повномасштабного українського електронного словника – **Інтегрованої лексикографічної системи «Словники України»**, словник поширюється на лазерних дисках, а також є у вільному доступі на створеному в Фонді Українському лінгвістичному порталі.

У Фонді створено та функціонує понад 40 **віртуальних лексикографічних лабораторій**, які в середовищі Інтернет забезпечують скоординовану роботу територіально розподілених дослідників над спільними проєктами (<http://lcorp.ulif.org.ua/LSlist/>).

Також у Фонді створено **паралельний українсько-російський лінгвістичний корпус** тексту Нового Завіту й здійснено його видання («Вічне Євангеліє»), розроблено теоретичні засади побудови концептографічної системи Святого Письма. Спільно з Науковою бібліотекою України імені В. І. Вернадського та Інститутом української мови НАН України Український мовно-інформаційний фонд НАН України здійснив два видання транслітерованого тексту визначної рукописної пам'ятки староукраїнської мови та мистецтва XVI століття «Пересопницького Євангелія» із науковим опрацюванням (впорядкування, дослідження, словопоказчики).

В УМІФ НАН України: відкрито лексикографічний ефект в інформаційних системах; побудовано інформаційно-структурну теорію лексикографічних систем (Л-систем); побудовано теорію інтегрованих, індексованих Л-систем; розроблено концепцію віртуальних систем професійної взаємодії (спільно з Інститутом кібернетики НАНУ), теорію та технологію віртуальних лексикографічних лабораторій; розроблено теорію лексикографічних

середовищ, яка стала теоретичним підґрунтям для побудови інтегрованих лексикографічних систем, зокрема ІЛС «Словники України»; розроблено теорію лінгвістичних онтологій. На цій основі із застосуванням корпусних технологій в онтологокерованих лексикографічних системах розроблено систему Трансдисциплінарні онтологічні діалоги об'єктно-орієнтовних систем (ТОДОС). Структурну схему ТОДОС наведено на рис.1.

Рис. 1. – Застосування корпусних технологій в онтологокерованих лексикографічних системах з використанням хмаринних та Grid-середовищ



Також в УМІФ НАН України розроблено теорію та технологію віртуалізації лексикографічних систем. На цій основі здійснено реалізацію УНЛК у хмаринному середовищі та на суперкомп'ютері СКІТ-3 (див. рис. 2).

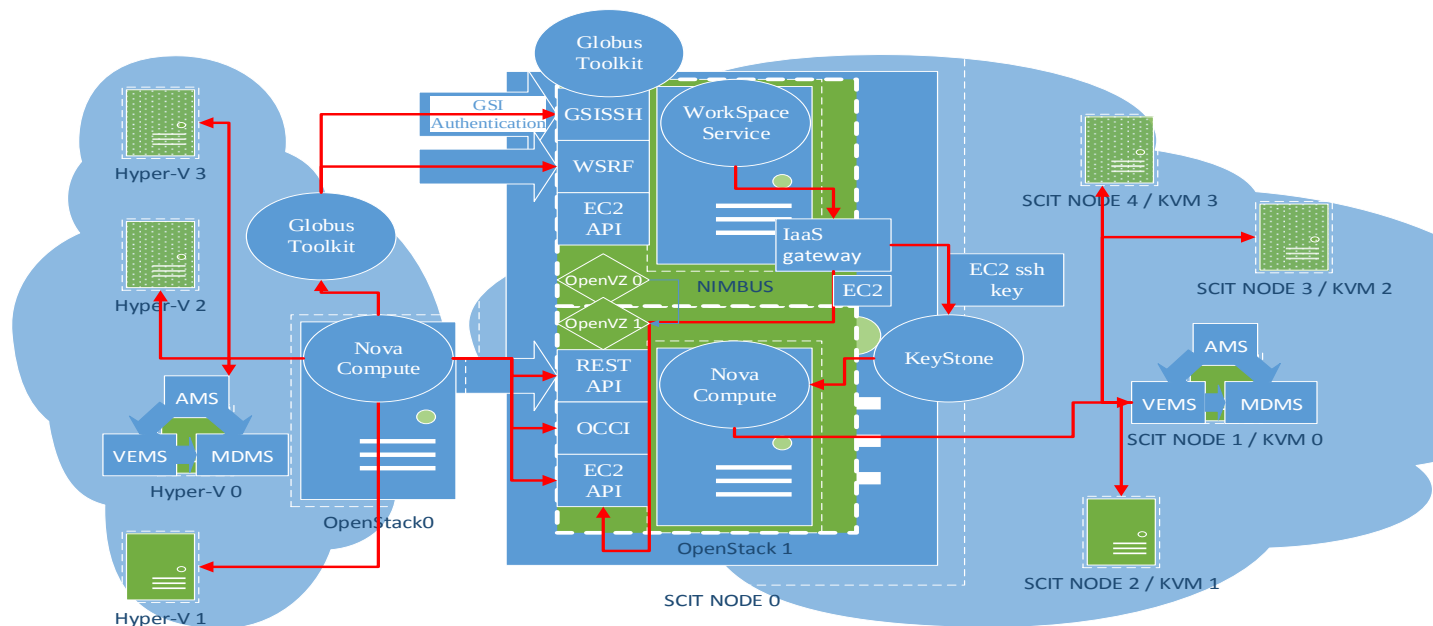


Рис. 2 – Використання суперкомп'ютерного комплексу SCIT для віртуалізації Л-систем.

Основні перспективні напрямки роботи Українського мовно-інформаційного фонду НАН України:

1. Впровадження нових апаратних систем (GPU, FPGA, біосистем тощо) у лексикографічну практику для прискорення обробки повнотекстової інформації.
2. Машинне навчання Л-систем.
3. **Створення системи лінгвістичної експертизи нормативно-правової бази України.** Зокрема, це стосується створення двох найважливіших елементів системи, а саме: Державного лінгвістичного корпусу нормативно-правової бази України та Державного тезаурусу України, що є першочерговими завданнями української держави, необхідними аспектами її цифрової трансформації (цифровізації).
4. Заснування на державному рівні системи громадської та наукової взаємодії – «Всеукраїнський лінгвістичний діалог» як розподіленої соціально-професійної мережі суб'єктів та об'єктів національного мовно-інформаційного процесу із включенням до нього сучасних віртуалізованих інструментів

лінгвістичного дослідження, лінгвістичних ресурсів та лінгвістичної інфраструктури.

5. Створення загальнонаціонального репозиторію національних інформаційно-лінгвістичних ресурсів «Всеукраїнського лінгвістичного діалогу» як ядра і центрального елемента національної лінгвістичної інфраструктури України.

Література

1. Надутенко Максим В. Віртуалізовані лексикографічні системи та їх застосування у прикладній лінгвістиці : дис. ... канд. техн. наук : 10.02.21 / НАН України, Нац. б-ка України ім. В. І. Вернадського. Київ, 2016. 284 с.
2. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков та ін. Т. 1 : Наукова парадигма та основні мовно-інформаційні структури. Київ. Український мовно-інформаційний фонд НАН України. 2018. 271 с.