неповною інформацією, відкриває великий простір для використання в задачах аналітики. Відповідно, оптимальні умови для використання таких способів наступні:

– великий обсяг бази (наявність неоднозначних слів та виключень);

– необхідність урахування смислового навантаження слова (в тому числі відокремлення омонімів та багатозначних слів);

– необхідність урахування зв’язків між словами (як-то смислові зв’язки, словотворення, структурні зв’язки);

До таких систем належать, зокрема, системи перевірки правопису, деякі семантичні мережі [9], пошукові системи та системи пошукової оптимізації.

Список використаних джерел

1. *Russom P.* Big data analytics, TDWI best practices report / Philip Russom. – Renton, WA: TDWI Research, 2011. – 35 с.
2. *Шнайер Б.* Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си / Брюс Шнайер. – Москва: Триумф, 2002. – 816 с.
3. PostgreSQL Documentation. Full Text Search. Dictionaries [Електронний ресурс] / The PostgreSQL Global Development Group. – 2015. – Режим доступу до ресурсу : http://www.postgresql.org/docs/9.1/static/textsearch-dictionaries.html.
4. *Henning K.* Performance analysis of MySQL’s FULLTEXT indexes and LIKE queries for full text search [Електронний ресурс] / Koch Henning. – 2013. – Режим доступу до ресурсу: http://makandracards.com/makandra/12813-performance-analysis-of-mysql-s-fulltext-indexes-and-like-queries-for-full-text-search.
5. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение / Андрей Анатольевич Зализняк. – Москва : Рус. яз., 1977.
6. *Smiley D.* Solr 1.4 Enterprise Search Server / D. Smiley, E. Pugh. – Birmingham : Packt Publishing Ltd, 2009. – 337 с.
7. Электронный словарь Мультитран [Електронний ресурс] – Режим доступу до ресурсу : http://www.multitran.ru/.
8. Collective Generation of Natural Image Descriptions / [P. Kuznetsova, V. Ordonez, B. Alexander C. et al.]. – Stony Brook, NY : Department of Computer Science, Stony Brook University, 2012.
9. WordNet, a lexical database of English [Електронний ресурс] – Режим доступу до ресурсу: https://wordnet.princeton.edu/.

УДК 004.82

Ю. І. Кисленко, А. В. Хіміч

ІНФОРМАЦІЙНА БАЗА ЛІНГВІСТИЧНОГО ПРОЦЕСОРА

Анотація: Стаття присвячена особливостям формування допоміжної інформаційної бази лінгвістичного процесора, орієнтованої на аналіз текстів. Основою дослідження постає підхід до структурного рівня мовної організації, що спирається на результати сучасних відомостей щодо мовленнєвої діяльності людини у багатьох суміжних напрямах та знімає багато суперечностей класичної лінгвістики.

Ключові слова: лінгвістичний процесор, індивідуальна мовна система, базова семантико-синтаксична структура, інформаційні природно-мовні технології, інформаційна база лінгвістичного процесора.

Вступ

Все більш нагальною постає проблема «спілкування» з комп’ютером природною мовою. Проте, прогрес у цій сфері поки що реалізується ще досить повільними темпами та не носить системного характеру. Звісно, все це зумовлюється складністю об’єкта дослідження та обмеженістю існуючих знань стосовно функціонування нашої нейромережі. Свідченнями такого стану є визнання розробників ІТ, орієнтованих на опрацювання природно-мовної інформації, стосовно двох ключових проблем у цій царині: відсутністю семантичного WEB та неспроможністю на сьогоднішній день формування потужних баз знань (природно-мовних), де б формувалася та відображувалася модель довкілля (а загалом, і модель світу), яка, власне, і зумовлює процес «розуміння» інформації. Таким чином, ці дві проблеми і визначають актуальність заявленої тематики розвитку цілого кластеру сучасних ІТ, орієнтованих на опрацювання природно-мовної інформації (ІПМТ).

На кафедрі технічної кібернетики НТУУ «КПІ» протягом останніх десятиріч формувався комплексний підхід до структурної організації мови, що базується на результатах сучасних досліджень мовленнєвої діяльності людини у багатьох помежованих напрямках, зокрема: біології, нейрофізіології, психології, філософії, кібернетики тощо. В результаті був сформований інтеграційний підхід до структурного рівня мовної організації, який знімає багато суперечностей класичної грамматики, розглянутих у праці [1] з міркуваннями щодо перспектив синтаксичних досліджень. Висновок з використання такого підходу є досить практичним: «Всі ми гарно та впевнено користуємося мовою (часто-густо не

знаючи як вона організована) лише за однієї умови – для всіх випадків актуалізації мовленнєвої діяльності на шляху синтезу чи аналізу мовного повідомлення нам достатньо лише однієї стандартної схеми – базової семантико-синтаксичної структури (БССС); і ця структура є похідною від структурно-функціонального рівня нейроорганізації зорового тракту». В концентрованому вигляді даний підхід презентований публікацією Ю. Кисленка «До витоків мовленнєвої діяльності» [2].

В цій праці ретельно проаналізована структура індивідуальної мовної системи (ІМС), введеної в обіг ще Л. Щербою [3], як сукупності лінгвістичного процесора (ЛП) та природно-мовної бази знань (БЗ). Лінгвістичний процесор, у загальному випадку, охоплює як всі знання стосовно мовної організації (словники, всі граматики тощо), так і сукупні знання (модель світу) стосовно нашого довкілля, що спресовані на мовному та образному рівнях в базі знань. Тож, для моделювання мовленнєвої діяльності нам потрібно сформувати як лінгвістичний процесор, відповідальний за аналіз та синтез мовного матеріалу, так і провести дослідження стосовно процедури формування моделі довкілля в БЗ, на якій відбувається інтерпретація (процес розуміння) довільного повідомлення. Стаття, власне, і присвячена проблемам моделювання лінгвістичного процесора на засадах запропонованого інтеграційного підходу до аналізу структурного рівня мовної організації.

1. Особливості системного підходу до структурного рівня мовної організації

Ключовою ідеєю досліджень у напряму усвідомлення та моделювання мовленнєвої діяльності постає концепція стосовно індивідуальної мовної системи (ІМС), зображеної на рис. 1, як сукупності компетенції особи стосовно мовної організації (лінгвістичний процессор – ЛП, за сучасними уподобаннями), так і всього потенціалу знань, накопичених на поточний момент в пам’яті (базі знань – БЗ) як на мовному, так і на сенсорному рівнях. У загальному випадку, мовленнєва діяльність актуалізується ІМС в режимах синтезу чи аналізу мовного повідомлення, що постають процесами одного порядку складності і мають отримати обґрунтування на рівні нейро-психофізіології.

Ідеї досліджень тісно пов’язані зі спробами формалізації наших знань стосовно реалізації лінгвістичного процесора – програмного продукту, відповідального за опрацювання природно-мовної інформації.

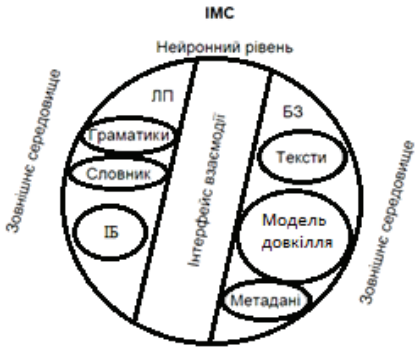

Рис. 1. ІМС

Основні принципові позиції системної організації, що формувалися на протязі багатьох років, презентовані працями «Архітектура мови» [4], «Від думки до знання» [5], «До витоків мовленнєвої діяльності» [2] та апробовані в лекційних кусах «Сенсорні системи», «Інформаційні природно-мовні технології». Базовими елементами для дослідження структурного рівня мовної організації постають поняття ***«ситуації»*** образного (зорового) рівня та БССС як вербалізованої форми відображення окремої ситуації, тобто фрагмента зорової складової довкілля, що потрапляє на центральну ямку сітківки та опрацьовується за повною програмою з визначенням всіх складових, їх прикмет та міри цих прикмет. Особливості організації зорового тракту, власне і зумовлюють кількість елементів ситуації (7 плюс/мінус 2), що одночасно ідентифікуються і в повній мірі опрацьовуються аналізатором.

Після опрацювання окремої ситуації спостерігачем вона закарбовується в його БЗ на образному рівні; а у випадку появи комунікативної інтенції (бажання чи необхідності поділитися з кимось своїми враженнями від баченого) реалізується процес трансляції окремої ситуації вже на мовний рівень.

Окрема ситуація транслюється на мовний рівень у вигляді окремої базової семантико-синтаксичної структури (БССС), де кожна складова ситуації образного рівня позначається відповідними мовними засобами. У вказаних працях подана інтерпретація БССС на змістовному, графічному та формальному рівнях. БССС – це двоскладова монопредикатна схема опису довільної ситуації реального чи віртуального світу, всі складові якої актуалізовані на атрибутивному рівні. Графічна інтерпретація БССС представлена на рис. 2. Таким чином, БССС, у загальному випадку, постає основним будівельним матеріалом формування довільного тексту, як на моно, так і на поліпредикатному рівнях.

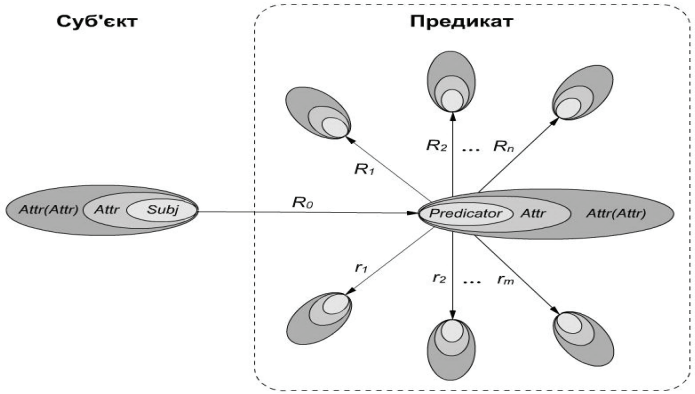

Рис. 2. БССС

2. Функціональний рівень організації лінгвістичного процесора

Лінгвістичний процесор, у загальному випадку, відповідальний за актуалізацію процесів опрацювання інформації. Матеріали даної статті, зокрема, пов’язані з процедурою аналізу мовного матеріалу; звичайно, аналізу автоматичного, пов’язаного з питаннями ідентифікації як окремих складових, так і всього повідомлення на моно та поліпредикатному рівнях його актуалізації. Саме для ідентифікації окремих фрагментів повідомлення у вигляді БССС, в структурі інформаційної бази і потрібна допоміжна інформація стосовно можливих варіантів подання окремих складових. Структурна організація інформаційної бази, практично і визначається особливостями формування повідомлень моно та поліпредикатного рівнів.

Попередні дослідження структурної організації текстів засвідчують, що множина повідомлень містить деякі елементи, що принципово не входять до структури БССС. Це обумовлюється тим, що суб’єктом, який формує та реалізує процедуру опису зовнішнього середовища виступає людина. Тому такі складові потрібно опрацьовувати на рівні інформаційної бази ЛП, враховуючи їх семантичне навантаження (рис. 3).

Аналіз текстів відбувається у декілька етапів. Спершу визначаються типи речень, з якими пов’язані певні мовні засоби. Другим кроком аналізу природномовних повідомлень постає виділення комунікативних складових, що визначають такі атрибути як ставлення до інформації, оцінку її вірогідності тощо. Ці складові визначають лише особливості комунікативного процесу, і не пов’язані з відтворенням конкретної ситуації.

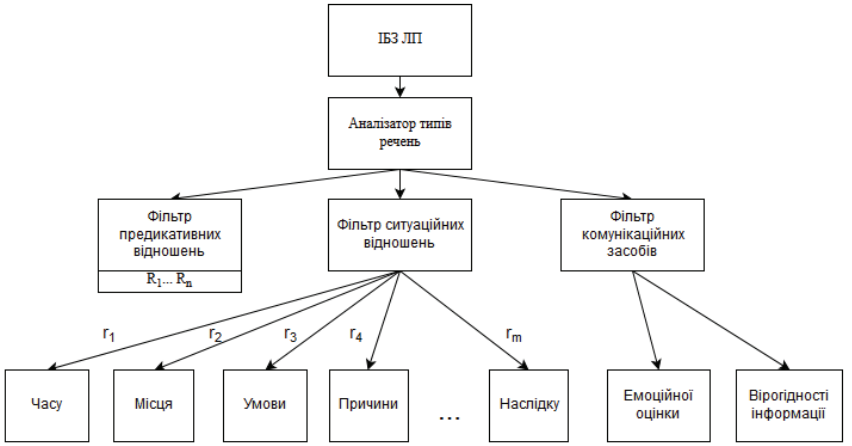

Рис. 3. Структура інформаційної бази ЛП

У табл. 1 представлено приклади мовних засобів, які для спрощення запису подано лише у безособовому вигляді. На першому рівні стоять характеристики комунікативного процесу, що стосуються джерела отримання інформації, другий рівень відтворює емоційну оцінку змістовного повідомлення, рівні 3, 4 та 5 визначають різні ступені вірогідності інформації [4]. Кожен з цих рівнів може бути деталізований, розширений та доповнений, наприклад, елементами етичних засобів, що ускладнює автоматичне заповнення БССС через наявність другорядної інформації.

Таблиця 1

Характеристики комунікативного процесу

№ п/п	Тип ставлення	Ставлення суб’єкта до повідомлення			
1	Джерело інформації	кажуть чути	за словами за даними	як кажуть як відомо	на погляд
2	Ставлення до змісту	вважається на щастя	на біду цікаво	важко зрозуміти	на нещастя
3	Достовірно	без сумніву безперечно	натурально зрозуміло	дійсно вірогідно	
4	Вірогідно	можливо	здається	мабуть	
5	Умовно	припустимо	нехай		

Предикативна складова БССС містить два типи відношень (предикативні та ситуаційні). Задача визначення цих відношень лягає на плечі лінгвістичного

процесора. Ситуаційні відношення можуть бути відносно легко ідентифіковані у більшості випадків через лексеми їхніх складових, що часто виступають у ролі обставин. Так, до ситуаційних відношень належать відношення, зазначені у табл. 2.

Таблиця 2

Ідентифікація ситуаційних відношень

Назва відношення	Питання, на які відповідає	Приклад
Способу дії	як? яким способом?	сидять **поруч**
Міри	якою мірою? як часто?	постукав **тричі**
Місця	де? куди? звідки?	стомилися **в дорозі**
Часу	як довго? з якого часу?	мовчав **кілька хвилин**
Причини	чому? з якої причини?	розсміявся **від радості**
Мети	навіщо? з якою метою?	зупинилися **на відпочинок**
Умови	за якої умови?	**за наявності квитка**
Наслідку	для якого наслідку?	взяв талісман **на щастя**

Часові та просторові відношення часто у тексті супроводжуються одиницями виміру відповідних величин, які в свою чергу поділяються на якісні та кількісні, як вказано в таблиці 3. [6] Слід відзначити таку специфіку використання одиниць виміру, як пов’язаність з мовними особливостями їх використання. Так, ідентифікація окремих проміжків часу та простору відбувається шляхом використання певної множини різномасштабних одиниць. При цьому система відношень організується таким чином, що послідовність одиниць є здатною з певною точністю презентувати проміжки часу чи простору. Це виступає допоміжною ознакою при ідентифікації ситуаційних відношень.

Будучи ідентифікованими, ситуаційні відношення дозволяють на поставити на своє місце цілу ситуацію, що визначає зв’язність тексту, обумовлену особливостями нашого мислення.

Таблиця 3

Одиниці виміру часу та простору

Кількісні одиниці		Якісні одиниці	
Час	секунда, хвилина, година, доба, тиждень, місяць, декада, квартал, рік, десятиріччя…	день, ніч, ранок, вечір…	час доби
		сніданок, обід…	час прийому їжі
Простір	метр, кілометр, сантиметр, міліметр…	понеділок, вівторок,…	послідовний порядок днів
		лікоть, сажень…	міри довжин

Варто відмітити що текстова інформація відтворює не весь безперервний простір, а лише окремі його дискретні складові: від кванта знань (як найменшої одиниці відтворення середовища), фрагмента (як сукупності квантів) до сукупності фрагментів (як достатнього рівня опису ситуації).

Висновки

Структурно-функціональний рівень організації ЛП – важлива складова індивідуальної мовної системи при опрацюванні природно-мовної інформації.

Проблема створення інформаційної бази лінгвістичного процесора постає важливим питанням на шляху до створення систем обробки природно-мовних текстів. Серед задач слід особливо відзначити важливість виділення відношень та комунікаційних елементів. Основний підхід для здійснення цього полягає у формуванні набору правил, за якими буде опрацьовуватися текст.

У роботі проаналізовано теоретичну основу для аналізу ситуаційних відношень лінгвістичним процесором та обґрунтовано необхідність первинної обробки тексту, що визначається як ітеративний процес, кінцевою ціллю якого є наповнення бази знань через ідентифікацію семантичних зв’язків. Також, тут представлено механізми визначення семантичних зв’язків ситуації та елементів, які не входять до складу БССС.

Список використаних джерел

1. *Астахова Л. И.* Предложение и его членение/прагматика, семантика, синтаксис / – Днепропетровский ГУ, 1992.
2. *Kyslenko Yuriy I.* Back to Basics of speech Activity, Biologically inspired Cognitive Architecture (2014) – № 8. – С. 47–69.
3. *Щерба Л. В.* О трояком аспекте языковых явлений и эксперименте в языкознании. // Языковая система и речевая деятельность. – М., 1974.
4. *Кисленко Ю. І.* Архітектура мови (Лінгвістичне забезпечення інтелектуальних інтегрованих систем), Навчальний посібник, 1998. – 343 с.
5. *Кисленко Ю. І.* Від думки до знання (нейрофізіологічне підгрунтя) : Монографія. – Київ, Видавництво «Український літопис», 2008. – 101 с.
6. *Кисленко Ю. І., Черевко О. С.* Категорії часу та простору в інформаційних природно-мовних технологіях // Адаптивні системи автоматичного управління – № 18, 2011. – С.62–70.