УДК 004.934007.811

Ю.І. Кисленко,Т.В. Кисленко

ЛІНГВІСТИЧНИЙ ПРОЦЕССОР В ІНФОРМАЦІЙНИХ ПРИРОДНОМОВНИХ ТЕХНОЛОГІЯХ

Анотація: Аналізується структурно-функціональний рівень організації лінгвістичного процесора як складової індивідуальної мовної системи, що постає ядром всіх сучасних інформаційних технологій, орієнтованих на опрацювання природно-мовної інформації. Особливість запропонованого підходу в тому, що структурний рівень мовної організації подається через множину однотипних базових семантико-синтаксичних структур, організованих за схемами моно чи полі предикатних рівнів.

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

Вступ

Інформаційні природно-мовні технології на даний час досягли вражаючих успіхів; створено безліч аналізаторів текстової та мовленнєвої інформації, які працюють, досить непогано, проте головний недолік їх у тому, що, практично, для довільного повідомлення вони намагаються кожен раз формувати унікальну синтаксичну структуру, починаючи з нуля цей непередбачуваний процес. В цьому контексті варто згадати, що для флективних мов одна структура простого речення з восьми компонентів може бути розгорнена десятками мільярдів варіантів. Цікаво також тут послатися на думку Перцова А.Н.[1] що розробниками ІТ вказаного напряму найчастіше постають математики, кібернетики, програмісти, а не лінгвісти. Тож тепер з'являється можливість систематизувати цей спорадичний процес та зорієнтувати його у цілеспрямоване прогнозоване русло.

Пропонується структурно-функціональна організація лінгвістичного процессора (ЛП), що спирається на засади інтегрального підходу до аналізу мовленнєвої діяльності людини, сформованого в останні десятиріччя на кафедрі ТК Національного технічного університету "КПІ". Матеріали такого підходу в концентрованому вигляді оприлюднені в статті "До витоків мовленнєвої діяльності" в журналі ВІСА-14 [2], постають плідною основою моделювання індивідуальної мовленнєвої системи (зі своїми складовими ЛП та БЗ) і визначають перспективні напрями розвитку цілого кластеру ІТ, орієнтованих на опрацювання природно-мовної (ПМ) інформації.

 ISSN 1560-8956

Платформа дослідження

Публікація в BICA-14 засвідчила знаковий етап формування авторської концепції структурної організації мови. З позицій інтегрування сучасних досліджень мовної діяльності у багатьох помежованих напрямах, включаючи: дослідників мовного ладу дитини, нейрофізіологів – дослідників зорового тракту, філософів, дослідників у сфері штучного інтелекту (використання онтогенетичних паралелей при дослідженні інтелектуальних процесів) і, звісно ж, лінгвістів вдалося синтезувати більш-менш несуперечливу концепцію становлення та розвитку мови у суспільстві. Наскільки сформований погляд автора стосовно мовної організації є адекватним реаліям мовної організації – час покаже. Проте отримані результати запропонованого підходу вже сьогодні дозволяють значним чином формалізувати шляхи формування цілого кластеру сучасних інформаційних технологій, орієнтованих на опрацювання природно-мовної інформації.

Основні досягнення запропонованого підходу, представлені вказаною публікацією, формують нову потужну платформу дослідження і моделювання мовленнєвої діяльності в рамках складових ЛП та БЗ індивідуальної мовленнєвої системи і можуть бути представлені наступними тезами.

1. За даними фізіологів та біологів чітко визначено поняття "ситуації" як фрагмента зорової складової довкілля, що попадає на центральну ямку сітківки та опрацьовується за повною програмою з ідентифікацією всіх складових *Obj*, *Subj* а також і дії Mov з визначенням всіх їхніх прикмет *Attr(Obj)*, *Attr(Subj)*, *Attr(Mov)* та і міри кожної з прикмет *Attr(Attr)*.
2. Трансляція окремої ситуації з образного рівня на мовний визначається у вигляді базової семантико-синтаксичної структури (БССС), що чітко визначена 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 (Рис.1).
3. За спостереженнями дослідників дитячої мови опанування мовою завжди відбувається при безпосередньому зіставленні конкретної ситуації довкілля з мовним її зіставленням.
4. Всі мовні категорії, незалежно від волі людини, опрацьовуються нашою нейромережею і з урахуванням інформації, отриманої не лише через зір, але й від усіх органів чуття.
5. Дослідження А. Гвоздєва [3] визначають головні етапи становлення БССС і дозволяють стверджувати, що дитина у віці 2.5 – 3 років опановує повністю структуру БССС як головну

схему відтворення мовними засобами довільної ситуації не лише реального, а й віртуального світу. Кажуть, що у цьому віці дитина стає "професором з лінгвістики" на рівні побутової мови. Послідовність цих етапів (Рис.2), з одного боку, визначає важливі стадії формування БССС як узагальненої схеми відтворення довкілля, а з іншого – подає перелік всіх можливих варіантів структур монопредикатного рівня, що формувалися на шляху онтогенезу, а отже – можуть і зустрічатися у мовному матеріалі. Саме ця послідовність і постає важливою на стадії формування лінгвістичного процесора.

6. Всі ми прекрасно володіємо мовою (найчастіше не знаючи як вона організована) лише за однієї умови – на всі випадки життя нам достатньо лише однієї структури БССС, якої достатньо для відтворення будь-яких життєвих колізій на моно або полі предикатному рівнях.
7. Таким чином, окрема ситуація (реального чи віртуального світу) постає складовою наших знань в БЗ на образному рівні, а окрема структура БССС – визначається квантом знань на символічному рівні.

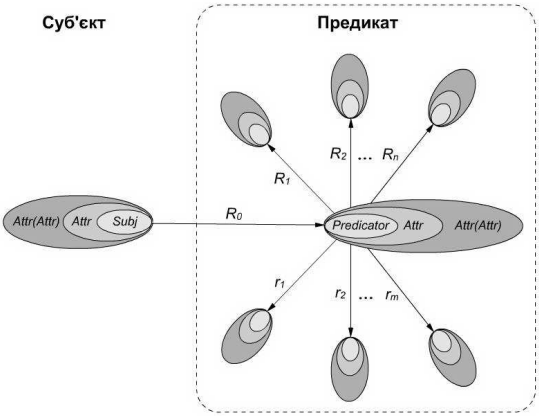

Рис. 1 – Базова семантико-синтаксична структура Subj - суб'єкт БССС, Predicator – дієслівне ядро предиката $P, R_1, \ldots, R_n$ – предикативні відношення, $r_1, \ldots, r_m$ – ситуаційні відношення R_0 – головне відношення "мати предикат"

Модульний принцип організації лінгвістичного процесора

Стрижнем досліджень у напряму усвідомлення та моделювання мовленнєвої діяльності постає концепція Л. Щерби [4] стосовно

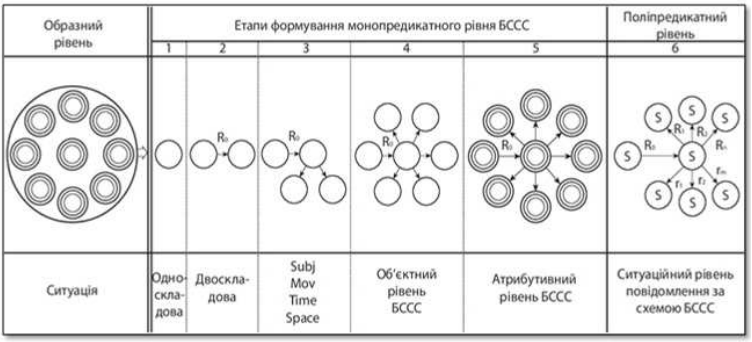

Рис. 2 – Головні етапи опанування дитиною мовного ладу

індивідуальної мовної системи (ІМС) як сукупності нашої компетенціїстосовно мовної організації (лінгвістичний процессор ЛП, за сучасними уподобаннями), так і всього потенціалу знань, накопичених на поточний момент в нашій пам'яті (базі знань БЗ) як на мовному, так і сенсорному рівнях. У загальному випадку, мовленнєва діяльність актуалізується ІМС в режимах синтезу чи аналізу мовного повідомлення, що постають "процесами одного порядку складності і не можуть стати зрозумілими без звернення до нейропсихо-фізіології".

Робота BICA14 [2] саме і присвячена, головним чином, дослідженню структурного рівня організації мовленнєвої діяльності, що, в свою чергу, власне і зумовлює структурно-функціональний рівень як бази знань, так і лінгвістичного процесора.

Пропоновані матеріали саме і пов'язані зі спробами формалізації наших знань стосовно реалізації лінгвістичного процесора – програмного продукту, відповідального за опрацювання природномовної інформації. Певні спроби працювати в цьому напряму реалізовувалися неодноразово і раніше [5,6], проте зараз з'явилася продуктивна платформа для їх продовження і систематизації.

Основою структурування ЛП постає послідовність етапів формування мовленнєвої діяльності на шляху онтогенезу/філогенезу мови, представлених рис. 2. Власне за висновками BICA14 [2], структурна організація мовленнєвої діяльності постає похідною від структурно-функціональної організації зорового тракту. Трансляція окремої ситуації на мовний рівень актуалізується базовою семантико-синтаксичною структурою.

Основні етапи формування БССС, представлені пунктами (1 - 5) рис. 2, власне визначають монопредикатний рівень формування повідомлення. Практично, поліпредикатний рівень також визначається особливостями організації БССС, тільки у цьому випадку окремими актантами постають не складові БССС, а конкре-

тні структури БССС (див. пункт 6, рис.1). Отже, головні особливості структурної організації довільного повідомлення визначаються вказаною схемою, не враховуючи лише особливості трансформування предикатора за схемами дієприкметникового, дієприслівникового зворотів та субстантива.

Маючи за взірець всі можливі варіанти структурної організації повідомлення, можемо вже досить цілеспрямовано підійти до проблем конструювання (формування) лінгвістичного процесора, орієнтованого на опрацювання довільного мовного повідомлення. Нагадаймо ще раз головний вердикт BICA-14 – будь яке повідомлення формується за схемами моно/полі предикатного рівнів, покриваючи схеми зворотів та рекурсивної організації повідомлення. Головна функція ЛП декомпозиція довільного повідомлення за окремими квантами знань, представленими стандартними структурами БССС, а БЗ – формування з множини квантів узагальненої моделі знань. Досягається ця мета за рахунок послідовного виконання певного переліку процедур. Тож, вже реальною постає проблема проектування ЛП на досить обґрунтованій формальній платформі.

Монопедикатний рівень організації ЛП

Коли основною функцією мови постає організація та подання знань (за Б.Ю. Городецьким [7]) а елементом сприйняття довкілля буде окрема ситуація реального чи віртуального світу з наступною її трансляцією на мовний рівень у вигляді структури БССС,то, мабуть, вузловим елементом ЛП повинен стати модуль ідентифікації БССС у всій своїй повноті як на атрибутивному, так і предикативному рівнях. Подамо далі основні напрями проектування ЛП.

1. Модуль фільтрації текстової інформації.

Перш ніж почати працювати з реальними текстами, необхідно зробити кілька важливих зауважень. Якщо головна функція мови – комунікативна, то існує широкий клас мовних засобів, які не входять до складу БССС як вербалізованої схеми відтворення довільної ситуації, але виконують важливу комунікативну функцію в побудові тексту, зачіпаючи певним чином і когнітивний рівень автора. Тому, перш за все, для роботи ЛП потрібен модуль ідентифікації таких засобів. В роботі [5] цілий розділ присвячений аналізу комунікативних засобів і подана їх класифікація. Ці дані повинні бути використані для наповнення БД лінгвістичного процесора. Тож, модуль ідентифікації комунікативних засобів повинен відфільтровувати, в першу чергу, засоби, що не входять потенційно до складу БССС. Інший клас подібних засобів – це ідіоматичні словосполучення, що функціонально теж не входять в базову структуру; їх також потрібно ввести до модуля БД лінгвістичного процесора.

 ISSN 1560-8956

2. Ідентифікація складових БССС на атрибутивному рівні.

Основними питаннями тут постають проблеми ідентифікації складових *Obj*/*Subj* та їх атрибутів *Attr(OBJ), Attr(Subj)* як узагальнених схем опису складових БССС (див. рис.3). Мабуть, основними проблемами тут постають питання формування флективного аналізатора з узгодженням флексій *Obj*/*Subj* з їхніми атрибутами, враховуючи, звісно, варіанти пре/пост позиції взаємного розташування.

На графічному рівні (рис. 3) опрацювання окремих складових ситуації образного рівня виглядатиме вже наступним чином. Спочатку зоровим аналізатором визначаються всі складники ситуації, далі ідентифікуються їх прикмети з великою роздільною здатністю і, нарешті, визначається міра кожної з прикмет: тобто формується весь атрибутивний рівень ідентифікації кожної складової.

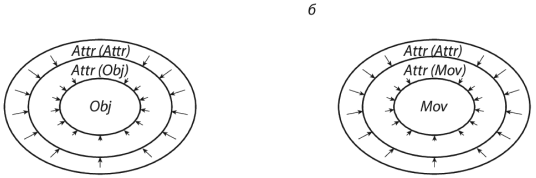

Рис. 3 – Опрацювання складових ситуації на образному рівні Obj – категорія об'єкта/суб'єкта , Mov – категорія руху (динамічна складова), Attr(Obj) – прикмети об'єкта, Attr(Mov) – прикмети руху, Attr(Attr) –міра прикмет

Ці схеми, практично, відтворюють структурну організацію складових ситуації *Obj, Subj* на множині їхніх прикмет *Attr(Obj*/*Subj)*. Важливо тут підкреслити графічну інтерпретацію структур *Obj*/*Subj*/*Mov*, бо на мовному рівні відсутні засоби ідентифікації їх взаємозв'язків, а лишаються прикметами лише семантичне навантаження та пре/пост позиція атрибутивних складових.

За подібною схемою пре/пост позиційного розташування ідентифікуються вже складові *Attr(Attr)*, що презентуються незмінними частинами мови – прислівниками. Майже аналогічним чином відбуватися має і процес ідентифікації предикатора БССС – Mov з послідовним визначенням його складових *Attr (Mov)* та *Attr(Attr)*, які презентуються прислівниками, знову ж таки з урахуванням пре/пост позиції їхнього розташування. Отже, цей етап ідентифікації окремих складових БССС закінчується ідентифікацією всіх її складників з урахуванням їх атрибутивного оточення.

3. Ідентифікація ситуаційних відношень БССС.

Наступним важливим етапом ідентифікації окремої структури постає розпізнавання ситуаційних відношень, що вказують на мі-

сце конкретної ситуації в рамках вже більш широкого контексту. В плані аналізу просторово-часових відношень автором були виконані свого часу дослідження, представлені роботами [3,4]. Слід зазначити, що цей модуль ЛП спирається на відповідну інформацію в БД стосовно одиниць виміру як часових, так і просторових відношень. Лишається тільки підкреслити, що багато в чому структурна організація просторових відношень співпадає з організацією часових, за виключенням хіба що одиниць виміру. В названих публікаціях досить ретельно проаналізовані особливості їх структурної організації. Слід лише нагадати, що ситуаційні відношення (сірконстанти) не обмежуються просторово-часовими відношеннями, а включають до свого складу обставини причини, наслідку, умови тощо, які теж потрібно ідентифікувати за певними прикметами.

4. Ідентифікація предикативних відношень.

Це, практично одне з найбільш проблемних питань ідентифікації октантів предиката, що у сукупності своїй формують модель керування дієслова. Авторам не відомі електронні версії моделей керування, хоча ряд фахівців займаються цими проблемами. Інший варіант формування моделей керування – накопичення статистик на початкових етапах функціонування ЛП.

Поліпредикатний рівень організації ЛП

Поліпредикатний рівень охоплює значну кількість структурних утворень на множині БССС. Можемо подати перелік окремих схем таких утворень, що мають, з нашого погляду, більш конструктивну та ширшу інтерпретацію порівняно з традиційним підходом.

1. Схеми ускладнення предикатора.

Структура БССС, як раніше вказувалося – вербалізована схема відтворення окремої динамічної ситуації з одним предикатом. Проте, ми часто маємо справу з більш складними динамічними ситуаціями, які принципово не можуть бути адекватно відтворені на рівні одного предикату з чітко визначеним ядром – предикатором. Сама вербалізація, що постає суто суб'єктивним процесом часто-густо використовує певні мовні засоби для відтворення таких категорій як: ставлення мовця до можливості виконання певної дії (модальні дієслова – можу, хочу, повинен, здатен …); засоби відтворення стадійності виконання дії (почати, продовжувати, закінчити тощо …); мовні засоби відтворення можливості опанування певними навичками (навчатися, опановувати, засвоювати …); і нарешті з'являється змістовне дієслово. У загальному випадку, кожна така характеристика принципово може відтворюватися окремим предикатом з усім своїм оточенням. В результаті матимемо цікаву структуру з одним суб'єктом Subj та множиною предикатів, що послідовно уточнюють один одного. Найчастіше, в реальних текстах, ми маємо комбінації таких мовних засобів з двох ком-

понент. Але, подамо узагальнену схему таких структур, що завжди можуть зустрічатися в реальних текстах, (див. рис. 4).

Особливість реалізації такої структурної схеми – в семантичному навантаженні всього ланцюга предикаторів, що не дозволяє змінювати порядок їх актуалізації, визначений автором. Окрім того, бачимо, що кожен наступний предикатор уточнює попередній в загальній динамічній картині динамічної ситуації, ініційованій одним загальним суб'єктом. За аналогічною схемою ускладнення предикатора формуються повідомлення поліпредикатного рівня, що знову ж таки пов'язані з системою прийняття рішень мовцем: "сподіваюся отримати…", "вирішую планувати…", "задумав почати працювати…" тощо, які, у загальному випадку, можуть поєднувати в межах окремого повідомлення кілька повноцінних (n-актантних) предикатів, ініційованих одним суб'єктом.

2. Схеми трансформування предикатора.

Статус ускладнених структур (не прості і не складні речення в класичній лінгвістиці) мають мовні конструкції поліпредикатного рівня, що пов'язані з трансформаціями предикатора за схемами дієприкметника, дієприслівника та субстантива. З нашої точки зору – це схеми поліпредикатного рівня організації повідомлення.

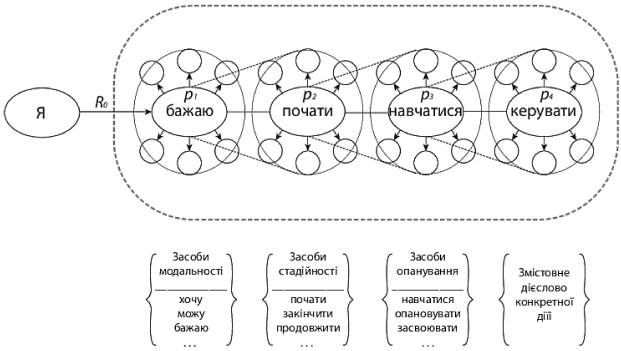

Рис. 4 – поліпредикатна схема ускладнення предикатора (Я дуже бажаю почати навчатися керувати авто)

Особливість організації таких структур в тому, що трансформований предикатор підпорядкованої структури може займати в кореневій структурі відповідні місця: об'єкта/суб'єкта *(Obj/subj)* у випадку субстантива, атрибутивного елемента *Attr(Obj/subj)* у випадку дієприкметника, та *Attr(Mov)* у випадку дієприслівника.

Особливість цих типів взаємодії двох предикатів у тому, що предикат підпорядкованої структури, практично у повному і незмінному вигляді входить на вакантне місце кореневої структури.

Це обов'язково важливо пам'ятати при опрацюванні текстової інформації – тобто потрібно враховувати, що при будь-яких трансформуваннях предикаторів цілісність окремої інкорпорованої структури БССС лишається непорушною. В якості прикладу на рис. 5 представлена структура повідомлення з дієприкметниковим зворотом.

3. Формування повідомлень підрядного рівня.

Монопредикатний рівень організації ЛП, практично, присвячений був мовним засобам формування окремої структури БССС, починаючи від об'єктного рівня окремих складових з їх атрибутивним оточенням та закінчуючи системою відношень предиката. Поліпредикатний рівень пов'язаний був з аналізом можливих схем ускладнення предикатора за поліпредикатною інтерпретацією зворотів. Поза увагою лишилися тільки схеми формування поліпредикатних повідомлень, що добре опрацьовані класичною лінгвістикою в межах складносурядних та складнопідрядних структур. Структура складносурядного зв'язку лишається поза нашою увагою. Важливо нагадати, що витоки підрядних схем взаємозв'язку все ж таки криються в множині функціональних зв'язків БССС, які усвідомлюються дитиною (людиною) вже на шляху філогенезу (див. пункт 6. рис. 2). Необхідно підкреслити, що функціонально зв'язки складових БССС визначені на трьох рівнях: *Obj*|*Subj*|*Mov, Attr (Obj*|*Subj*|*Mov)* та міри цих атрибутів *Attr(Attr)*. Поліпредикатна схема зв'язків базується лише на складових *Obj*|*Subj*|*Mov* та їх прикметах.

Реалізується такий зв'язок, з одного боку, за допомогою парних мовних засобів "той . . . ., хто", "того . . . кого", що ідентифікують функціональне навантаження конкретної складової; наприклад "Той, хто греблю рве на волі". Інша схема формування поліпредикатного рівня пов'язана вже з атрибутивним рівнем і ідентифікаторами такого зв'язку постають вже представники знову ж таки парних мовних засобів "такий, як. . .", "такого, що. . .", "студент, який. . ." тощо.

Висновки

З позицій інтегрального підходу до аналізу структурного рівня мовної організації, що спирається на формальне визначення ситуації, яка транслюється на мовний рівень у вигляді структури БССС, проаналізовано структурний рівень мовної організації. В роботі ретельно проаналізовані особливості структурної організації мовного матеріалу, реалізованого за схемами моно/полі предикатного рівнів. Взаємодія різних структурних варіантів формування повідомлення забезпечує формування повідомлень довільного рівня складності, виходячи на рівень рекурсивної організації мовного матеріалу. Головна функція лінгвістичного процесора – це декомпозиція повідомлення за базовими структурами БССС з визначенням

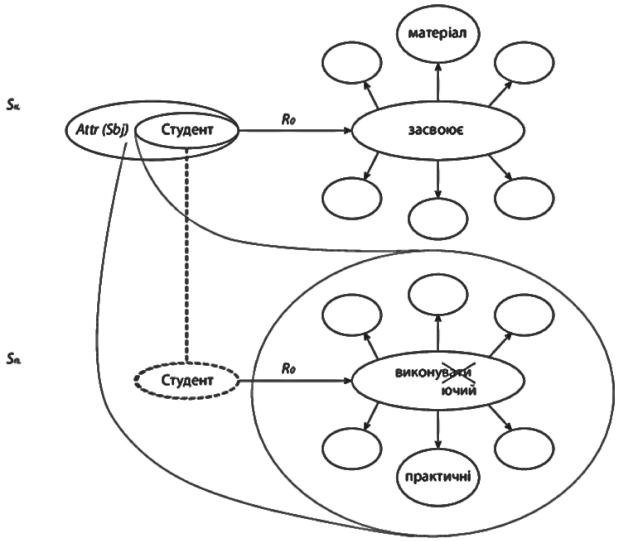

Рис. 5 – Поліпредикатна інтерпретація дієприкметникового звороту (Студент, виконуючий практичні завдання, краще засвоює матеріал.)

функціонального навантаження всіх складових. Багаторічна практика аналізу довільних текстів (прози та поезії) засвідчує саме монізм структури БССС в плані структурної організації як окремого повідомлення так і мови, загалом, що відкриває обнадійливі перспективи у моделюванні індивідуальної мовної системи як основи розвитку цілого кластеру інформаційних природно-мовних технологій.

Список використаних джерел

1. Перцов Н.В. О некоторых проблемах современной семантики и компьютерной лингвистики // Московский лингвистический альманах, 1996, - Вып. 1, с. 9-66.
2. Kislenko Yuriy I. Back to Basics of speech Activity, Biologically inspired Cognitive Architecture (2014) 8, 47 69.
3. Гвоздев А.Н. Формирование у ребенка грамматического строя русского языка: – М.: Изд-во АПН, 1949.
4. Щерба Л.В. О трояком аспекте языковых явлений и эксперименте в языкознании. // Языковая система и речевая деятельность. – М., 1974.

5. Кисленко Ю.І. Архітектура мови (Лінгвістичне забезпечення інтелектуальних інтегрованих систем), Навчальний посібник, 1998, 343с.

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.

7. Городецький Б.Ю. Компьютерная лингвистика: моделирование языкового общения. (Пер. с англ. /сост., ред. и вступ. ст. Б.Ю. Городецкого). - Серия “Новое в зарубежной лингвистике”, Вып 24. – М.: Прогресс, 1989.- 432 с.

Отримано 15.10.2015 р.