МІРАМ Г. Е.
Київський національний університет імені Тараса Шевченка

МОДЕЛЮВАННЯ МОВНИХ ОДИНИЦЬ В ЛІНГВІСТИЦІ ТА ІНФОРМАТИЦІ: З ІСТОРІЇ ПИТАННЯ

Розглядаються у ретроспективі підходи і методи конструювання лінгвістичних моделей і моделей інформатики. Робиться загальний висновок про те, що основними проблемами лінгвістичного моделювання є еклектичність підходів і суб'єктивність визначення первинних синтаксичних і семантичних параметрів, чим пояснюються відомі недоліки систем автоматичної обробки текстів.
Ключові слова: лінгвістичне моделювання, моделі інформатики, автоматична обробка текстів.

Рассматриваются в ретроспективе подходы и методы создания лингвистических моделей и моделей информатики. Делается общий вывод о том, что основными проблемами лингвистического моделирования являются эклектичность подходов и субъективность присвоения исходных синтаксических и семантических параметров, что определяет известные недостатки систем автоматической обработки текстов.
Ключевые слова: лингвистическое моделирование, модели информатики, автоматическая обработка текстов.

Methods of and approaches to linguistic and IT modeling are reviewed in retrospect. Eclectic approach and subjective definition of initial syntactic and semantic parameters are suggested as the main source of problems leading to the known drawbacks of automatic text processing systems.
Key words: linguistic modeling, IT models, automatic text processing

Завдання побудови синтаксичних і семантичних класифікацій мовних одиниць, яке ще на початку минулого століття могло вважатись суто теоретичним, в наші часи стає практичною проблемою, від вирішення якої у певній мірі залежить успішність функціонування систем автоматичної переробки тексту різного призначення, від машинного перекладу до систем автоматичного пошуку інформації.

Проте створення подібної класифікації в наші часи, як і майже століття тому назад, залежить від певного комплексу іманентних мовних якостей, що обумовлюють точність і відтворюваність моделі, тому ретроспективний аналіз підходів і методів моделювання, якому присвячується ця робота, залишається **актуальним** і на сучасному етапі.

Таким чином, **предметом** даного дослідження є ретроспективний аналіз лінгвістичного моделювання, **об'єктами**

якого слугують методи і підходи у створенні таких моделей в восьмидесяті – дев'яності роки минулого століття.

Як раніше, так і тепер основним завданням тих, хто створює суто лінгвістичну модель або так званий мовний інтерфейс для інформаційної системи [1; 2] є "… вилучення з мови її концептуального змісту і надання йому такої форми, що дозволила б ввести його в комп'ютер для наступної обробки …" [3:7][3].

Однак незважаючи на спільність завдання, підходи до моделювання суттєво відрізняються один від одного.

Перш за все, це стосується розбіжності у підходах між традиційним мовознавством, що дотримується інтуїтивних методів мовної класифікації [4; 5; 6], і структурною лінгвістикою, що, описуючи мовні явища, виходить з форми і в своїх екстремальних проявах взагалі заперечує значення (див., наприклад, роботи [7; 8]).

Слід відмітити, однак, що поступово підходи традиційної лінгвістики і структуралізму починають наближатись один до одного – традиційна школа усвідомлює необхідність об'єктивізації, а структуралісти відмовляються від абсолютизації форми, розуміючи що "… мова – це система, яка дуже складним чином реалізує посередництво між світом значення й світом звуку" [9 : 27].

Проте розбіжності між підходами залишаються. Спробуємо їх узагальнити на прикладі класифікації частин мови [10].

Існують наступні основні підходи до створення такої класифікації.

(1) ***Семантичний***. Встановлення кореляції між значенням і приналежністю до певної частини мови.

(2) ***Морфологічний***. Частини мови виділяються на базі особливостей словотворення та словозмінення. Такий підхід є досить ефективним для флективних мов з розвиненою морфологією[4], але непридатним для ізолюючих мов (наприклад, китайської) і мов з високим ступенем аналітизму.

(3) ***Синтаксичний***. Для виділення частин мови використовують функціональний (синтаксичний) критерій. В

[3] Переклад мій (Г. М.)

[4] Морфологічна база класифікації є основною причиною успішності алгоритмів аналізу при машинному перекладі з російської [12, 13, 14]

екстремальному варіанті виділені таким чином частини мови співпадають з членами речення.

(4) ***Психолінгвістичний***, тобто такій, що базується на інтуїції. Вважається, що останній є найбільш розповсюдженим. Крім того, інтуїція лінгвістів-носіїв мови відіграє важливу роль в кожному з попередніх підходів.

Розмаїття підходів до однієї лише класифікаційної проблеми ілюструє складність загальної проблеми лінгвістичного моделювання, особливо, у світлі застосування моделей у прикладних системах, де найважливіші характеристики – це об'єктивність і відтворюваність [11].

В роботах Л. В. Щерби [15; 16], Г. О. Винокура [17], І. І. Мєщанинова [18] та інших представників російської граматичної школи відстоюється пріоритет лінгвістичної інтуїції. Проте не можна ігнорувати той факт, що багато інтуїтивно загально визнаних фактів були спростовані перевіркою на інженерно-лінгвістичних моделях [19; 20] та результатами машинних статистичних досліджень [21].

Крім того, широке визнання дихотомії "мова-мовлення" виявило ще одну складність у підходах до моделювання, пов'язану з тим, яка одиниця має бути об'єктом моделювання – одиниця мови, чи одиниця мовлення.

Розвиток текстології у восьмидесяті-дев'яності роки (див., наприклад, [23; 24; 25; 26] і щільна увага до текстових параметрів в інформатиці [1; 27] призвели до спроб розмежування об'єктів моделювання (об'єкти мови та об'єкти мовлення). Нажаль, ці спроби не можна назвати вдалими навіть на сучасному етапі, і в більшості моделей первинними об'єктами служать одиниці мови.

Разом з тим в первинних параметрах (класифікаціях базових одиниць) інженерно-лінгвістичних моделей текстовим змінним, перш за усе, відношенням "тема-рема" стали приділяти більше уваги [28; 29; 30].

Таким чином, узагальнюючі, можна стверджувати, що навіть цей побіжний огляд різних підходів до створення лінгвістичних класифікацій і моделей наочно демонструє той факт, що всі розбіжності зводяться до двох ключових питань: ступеня інтуїтивного і врахування/неврахування текстових параметрів.

На наш погляд основу лінгвістичних класифікацій (морфологічну, синтаксичну чи семантичну) не слід вважати критичною характеристикою моделі, оскільки вибір класифікаційної основи часто залежить від будови мови. Крім того, в логічно побудованій моделі основи мають взаємно перетворюватись одна в одну [31] і основа класифікації має значення лише як засіб забезпечення більшої об'єктивності на первинному етапі.

Варто зупинитись тепер на розбіжностях між моделями лінгвістики і моделями інформатики. Е. Попов [27] пропонує оцінювати ті та інші моделі за наступними критеріями.

1. Ступень охоплення основних явищ процесу спілкування : а) мова чи мовлення; б) дискурс або речення; в) зміст чи псевдозміст; г) врахування лакун; д) врахування синонімії і омонімії; е) роль екстралінгвістичних знань.

2. Обсяг знань, що характеризують глибину розуміння тексту (модель мови, модель екстралінгвістичного оточення, модель учасників спілкування).

Серед наведених критеріїв найважливіші для розмежування моделей лінгвістики та інформатики, на нашу думку, є застосування в останніх текстових параметрів і екстралінгвістичних знань.

Слід відмітити, однак, що вищенаведені головні критерії справедливі лише для ідеального випадку, в реальності ж при створенні моделей спостерігається еклектичність підходів.

Слід відмітити також, що відношення до текстових параметрів теж не є однаковим.

Домінує модель мови, але деякі лінгвісти наполягають на необхідності побудови як моделі мови, так і моделі мовлення [32]. В теорії та практиці машинного перекладу все більша увага приділяється текстовим параметрам [20; 30; 33]. Наголошується також необхідність мати "базу знань" у системах машинного перекладу [34; 35].

Для кращого розуміння ситуації розглянемо концепції, що складають підґрунтя основних синтаксичних і семантичних моделей.

Однією з найбільш розповсюджених методик семантичного моделювання є компонентний аналіз [36; 37; 38], хоча цей метод ширше застосовується у дослідженнях, ніж у прикладних моделях.

Цей метод виходить з припущення, що семантику мовної одиниці можна описати через зіставлення з певними універсальними ознаками (семами), такими як “стать”, “категорія виміру”, “категорія істоти”, тощо. Результатом аналізу (моделлю) слугує набір сем.

Деякі дослідники вважають, що модель компонентного аналізу занадто громіздка, однак, на нашу думку, основним недоліком моделі є суб’єктивність набору сем і методик зіставлення, що негативно впливає на відтворення результатів. Складність і ємність структури в наші часи вже не є перешкодою у практичному застосуванні моделі, зважаючи на майже необмежені можливості сучасних комп’ютерів.

Концепцію “семантичних відмінків” Ч. Филмора [39; 40] можна розглядати як подальший розвиток компонентного аналізу стосовно до текстових структур.

Виходячи з відомих ідей семантичних множників і предикатів формальної логіки, пропонується не тільки обчислювати аргументи предикатів – агент, контрагент, об’єкт, атрибут, суб’єкт, інструмент і т. д, але і вказувати їх “семантичні відмінки” (ролі). Семантичні відмінки використовуються у багатьох прикладних розробках [див., наприклад, 41; 42].

Самим суттєвим обмеженням цієї моделі, як і попередніх, є суб’єктивність, як самого набора ролей, так і присвоєння їх аргументам предикату, що призводить до неоднозначності у первинних параметрах моделі.

Варто також звернути увагу на запропоноване Ч. Филмором виділення у рамках мовного значення так званої “пресу позиції”, тобто екстралінгвістичних знань, що у деяких випадках не можна віднести ані до дефініції, ані до сполучуваності, і які суттєво впливають на зміст [43].

Інженерно-лінгвістичні моделі демонструють еклектичність застосованих концепцій семантичного аналізу – не існує розмежування між значенням у мові, значенням у мовленні (змістом) і екстралінгвістичним значенням, що часто призводить до невірних висновків [44; 45; 46].

Еклектичність проявляє себе і у синтаксичних моделях (див., наприклад, [47]). Крім того невирішеність і на даний час проблеми

синтаксичного аналізу свідчить про відсутність нових працездатних гіпотез про синтаксичну структуру речення [48; 49].

Достатньо показовими в цьому сенсі є моделі "семантики преференції" [32; 50] і "концептуальної залежності" [51].

Моделі "семантики преференції" базуються на двох семантичних конструктах – на "формулі слова", що виражає значення через елементарні семантичні ознаки, і на "шаблоні" – своєрідній логіко-семантичній моделі повідомлення. В процесі аналізу текст розділяють на фрагменти за синтаксичними маркерами і алгоритм виконує зіставлення "формул слів" у фрагменті з "шаблонами" і будує семантико-синтаксичне зображення тексту фрагмента.

Модель "концептуальної залежності" за своєю ідеологією нагадує "семантики преференції". Як і попередня, вона відноситься до так званих "асинтаксичних" моделей, де ігноруються речення і аналіз відбувається за текстовими фрагментами. Модель працює з семантичним зображенням у вигляді "мережі концептуальних залежностей". Основне завдання моделі отримання певного семантичного інваріанта на базі лінгвістичної і екстралінгвістичної інформації і виведення з нього характеристик і обставин закодованої у тесті дії.

Типовою рисою моделі "концептуальної залежності" є еклектичність – однаковий статус лінгвістичних і енциклопедичних знань, одиниць мови і мовлення, тощо.

Найскладнішою за своїми синтаксичними і семантичними конструктами моделлю можна вважати модель "Зміст-Текст" [52; 53; 54]. Крім того, модель "Зміст-Текст" відрізняє послідовність у дотриманні суто лінгвістичних підходів і одиниць: енциклопедичні знання повністю виключаються, модель орієнтована на речення і має розвинений синтаксичний і морфологічний компонент.

Наприкінці восьмидесятих – дев'яностих років в прикладних інженерно-лінгвістичних системах (особливо закордонних) типовим стає застосування так званої "уніфікованої граматики" – алгоритмічного комплексу аналізу – синтезу синтаксичних / семантичних параметрів на базі уніфікованих семантико-синтаксичних характеристик (так званих "тегів") (див., наприклад, [47; 55; 56]).

Проте суттєво ситуацію поява такого уніфікованого комплексу не змінила – як і раніше основними проблемами лінгвістичного моделювання залишаються еклектичність підходів і суб'єктивність визначення первинних синтаксичних і семантичних параметрів, чим здебільшого і пояснюються відомі недоліки різних систем автоматичної обробки текстів – машинного перекладу, інформаційного пошуку тощо.

ЛІТЕРАТУРА

1. *Попов Э. И.* Общение с ЭВМ на естественном языке / Эдуард Викторович Попов. – М. : Наука, 1982. – 360 с. 2. *Поспелов Д. А*. О "человеческих" рассуждениях в интеллектуальных системах // Логика рассуждений и ее моделирование / Дмитрий Александрович Поспелов. – М. : Научный совет по комплексной проблеме "Кибернетика" АН СССР, 1983. – С. 5–37. 3. *Звегинцев В. А.* Язык и лингвистическая теорія / Владимир Андреевич Звегинцев. – М. : Эдиториал УРСС, 2001. – 248 с. 4. *Виноградов В. В.* Основные типы лексических значений слова / Виктор Владимирович Виноградов // Избранные труды. Лексикология и лексикография. – М. : Наука,1977. – С. 162–189. 5. *Винокур* Г. О. Заметки по русскому словообразованию / Григорий Осипович Винокур // Избранные работы по русскому языку. – М. : Наука, 1959. – 451 с. 6. *Мещанинов И. И.* Члены предложения и части речи / Иван Иванович Мещанинов. – Л. ; М. : АН СССР, 1945. – 319 с. 7.*Bloomfield L.* Language / Leonard Bloomfield. – University of Chicago Press, 1933. – 564 с. 8. *Harris Z. S.* Methods in structural linguistic / Z. S. Harris – Chicago University Press. 1951. – 187 p. 9. *Чейф У. Л.* Значение и структура языка / Уоллес Л. Чейф ; [пер. с англ.]. – М. : Прогресс, 1975. – Серия : "Языковеды мира". – 432 с. 10. *Алпатов В. М.* О разных подходах к выделению частей речи // Вопросы языкознания / Владимир Михайлович Алпатов. – 1986. – № 4. – С. 37–46. 11. *Пиотровский Р. Г.* Математическая лингвистика / [Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская]. – М. : Высшая школа, 1977. – 383 с. 12. *Hiroaki K.* Speech-to-speech translation : A Massively Parallel Memory-Based Approach / Hiroaki Kitano // Computational Linguistics. – 1994. – Vol. 21, № 4. – Boston : Kluwer Academic Publishers. – P. 590–592. 13. *Pigot I.* SYSTRAN Maschinenübersetzung bei der Kommission der EG Gegenwärtiger Stand der Geschichte / Ian Pigo // Linze Linguistic server. – Berlin, 1985. – S. 22–27. 14. *Podiumsdiskussion* Maschinelle Übersetzung im Aufwind // Linze Linguistic server. – 1985. – S. 58–65. 15. *Щерба Л. В.* Языковая система и речевая деятельность / Лев Владимирович Щерба. – М. : УРСС Эдиториал, 2004. – 432 с. 16. *Щерба Л. В*. Избранные работы по русскому языку / Лев Владимирович Щерба. – М. :Учпедгиз, 1957. – 187 с. 17. *Винокур Г. О.* Избранные работы по русскому языку / Григорий Осипович Винокур. – М. : Учпедгиз, 1959. – 492 с. 18. *Мещанинов И. И.* Предикативность, сказуемость, глагольность / Иван Иванович Мещанинов // Вестник ЛГУ. – 1946. – Т. 4–5. – С. 119–132. 19. *Новиков А. И.* Семантика текста и ее формализация / Анатолий Иванович Новиков. – М. : Наука, 1983. – 215 с. 20. *Марчук Ю. Н.* Методы моделирования перевода / Юрий Николаевич Марчук. – М. : Наука, 1985. – 199 с. 21. *Brown et al* A Statistical Approach to Machine Translation // Computational Linguistics. – 1990. – V. 16,2. – С. 79–85. 22. *Бондарко А. В.* Значение и смысл: проблема интенциональности / Александр Владимирович Бондарко // Лингвистика на исходе XX века: итоги и перспективы / ред. И. М. Кобозева. – М. : МГУ им. М. В. Ломоносова, 1995. – С. 68–69. 23. *Ван Т. Дейк* Вопросы прагматики текста // Новое в зарубежной лингвистике / Т. Ван Дейк. – М. : Прогресс, 1978. – Т. 8. – 336 с. 24. *Колшанский Г. В.* Логика и структура языка / Геннадий Владимирович Колшанский. – М. : Высшая школа, 1965. – 240 с. 25. *Латышев Л. К.* Перевод: проблемы теории, практики и методики преподавания / Лев Константинович Латышев. – М. : Просвещение, 1988. – 160 с.

26. *Леонтьева Н. Н.* Общесемантический компонент в системе понимания текста / Нина Николаевна Леонтьева // Проблемы прикладной лингвистики. – М. : Прогресс, 2001. – С. 18–20. 27. *Попов Э. В.* Общение с ЭВМ на естественном языке / Эдуард Викторович Попов. – М. : Наука, 1982. – 360 с. 28. *Слюсарева Н. А.* Проблемы функциональной морфологии современного английского язика / Наталья Александровна Слюсарева. – М. : Наука, 1986. – 216 с. 29. *Нелюбин Л. Л.* Учебник военного перевода / [Л. Л. Нелюбин, А. А. Дормидонтов, А. А. Васильченко]. – М. : Воениздат, 1981. – 379 с. 30. *Нелюбин Л. Л.* Перевод и прикладная лингвистика / Лев Львович Нелюбин. – М. : Высшая школа, 1983. – 207 с. 31. *Ревзин И. И.* Модели языка / Исаак Иосифович Ревзин. – М. : Изд-во АН СССР, 1962. – 191 с. 32. *Wilks Y. A.* An intelligent analyser and understander of English / Yorick Wilks // CACM, 1975. – V. 18, N. 5. – P. 264–274. 33. *MacRoy S.* Using multiple knowledge sources for word sense discrimination", Computational Linguistics, 18(1) / Susan McRoy, 1992. – P.1–30. 34. *Batori* Papadigmen in maschineller Sprachuebersetzung – Neue Ansatze in Maschineller Sprachbearbeitung. – Tuebungen, 1986. 35. *Hutchins W.* Machine Translation: Past, Present, Future. Ellis-Horwood Limited, Chichester, England / W. Hutchins, 1986. 36. *Комлев М. Г.* Компоненты содержательной структуры слова / Николай Георгиевич Комлев. – М. : Наука, 1969. – 192 с. 37. *Медникова Э. М.* Значение слова и метод его описания / Эсфирь Максимовна Медникова. – М. : Высшая школа, 1974. – 202 с. 38. *Смолина К. П.* Компонентный анализ и семантическая реконструкция в истории слова / К. П. Смолина // Вопросы языкознания. – 1986. – № 4. – С. 97–104. 39. *Филмор Ч.* Основные проблемы лексической семантики / Ч. Филмор // Новое в зарубежной лингвистике. – М. : Прогресс, 1983. – Вып. 12. – С. 74–122. 40. *Katz J. J.* An Integrated Theory of Linguistic Description. – Cambridge, The M.l.T. Press, 1964. 41. *Нагао М.* Машинный перевод с японского языка на английский / [Нагао М., Цудзии Дз., Накамура Дз.] – ТИИЭР, 1986, – Т. 74, № 7. – С. 112–133. 42. *Мирам Г. Э.* Система машинного перевода "СИМПАР" – принципы разработки и задачи экспериментальной эксплуатации / Г.Э.Мирам, О.Н.Гальченко // Проблемы автоматического и экспериментально-фонетического анализа текстов. – Минск, 1986. – С. 141–146. 43. *Кифер Ф.* О пресуппозициях / Ф. Кифер // Новое в зарубежной лингвистике. – 1978. – Вып. VIII. – С. 337–353. 44. *Конецкая В. П.* Социология коммуникаций / В. П. Конецкая. – М. : Международный университет бизнеса и управления, 1997. – 304 с. 45. *Сидоров Е. В.* Системная модель коммуникации и параметры текста в переводе / Е. В. Сидоров // Лингвистические проблемы перевода. – М. : Изд-во МГУ, 1981. – С. 43–53. 46. *Цибова И. А.* Определите значение слова / И. А. Цибова. – М. : Изд-во Международные отношения, 1981. – 192 с. 47. *Miram G. E.* Translation Algorithms / Guennadi Miram. – К. : Твім-Інтер, 1998. – 175 с. 48. *Кошевая И. Г.* Текстообразующие структуры языка и речи / И. Г. Кошевая. – М. : МГПИ им. В. И. Ленина, 1983. – 181 с. 49. *Рахманкулова И. С.* Моделирование предложений и семантика глаголов: Учебное пособие по курсу теоретической грамматики немецкого языка для студентов 3–5 курсов факультета немецкого языка / Изюм-Эрик Салиховна Рахманкулова. – М. : МГПИИЯ, 1973. – 101 с. 50. *Wilks Y.* Semantics and world knowledge in Machine Translation / Yorick Alexander Wilks // FBIS Seminar on Machine Translation, (Rosslyn Virginia, 8–9 March 1976). – 1976. – P. 67–69. 51. *Schank R.* Conceptual Dependency: Theory of Natural Language Understanding / Roger C. Schank // Cognitive Psychology, 1972. – С. 532–631. 52. *Мельчук И. А.* Опыт теории лингвистических моделей "Смысл-Текст"/ И. А. Мельчук. – М. : Школа "Языки русской культуры", 1999. – 346 с. 53. *Апресян Ю. Д.* Об идеологии системы ЭТАП-2 / Юрий Дереникович Апресян, Леонид Львович Цинман // Формальное представление лингвистической информации. – Новосибирск, 1982. – С. 3–19. 54. *Марчук Ю. Н.* Проблемы машинного перевода / Юрий Николаевич Марчук. – М. : Наука, 1983. – 234 с. 55. *Shieber S.* An Introduction to Unification-Based Approaches to Grammar / Stuart M. Shieber. – V. 4 of CSLI Lecture Notes Series: Center for the Study of Language and Information, Stanford, CA, 1986. 56. *Pulman S. G.* Unification encodings of grammatical notations / Stephen G. Pulman // Computational Linguistics. – Vol. 22, 1996. – P. 295–328.