УДК (038):[655.5:681.3]

***І. С. Кравчук***

*Харківський національний університет імені В. Н. Каразіна, м. Харків*

# Лінгвістичні метамоделі

**Кравчук І. С. Лінгвістичні метамоделі.** Стаття присвячена методам оцінки альтернативних лінгвістичних моделей різних типів. Систему критеріїв такої оцінки для вибору однієї з конкуруючих об'єктних моделей названо мета моделлю. Розглянуто різні типи мета моделей у залежності від різних типів об'єктних моделей. Розмежовано два типи моделей: породжуючі й перетворюючі. Уточнено критерій повноти для великих баз даних. Уведено два критерії внутрішньої довершеності моделей: дескриптивна й процедурна простота. Запропоновано кількісні міри якості для породжуючих і перетворюючих моделей.

***Ключові слова: об'єктні моделі, мета моделі, декларативна простота, процедурна простота, відстань між текстами.***

**Кравчук И. С. Лингвистические метамодели.** Статья посвящена методам оценки альтернативных лингвистических моделей разных типов. Система критериев такой оценки для выбора одной из конкурирующих объектных моделей названа метамоделью. В статье рассмотрены разные типы метамоделей в зависимости от разных типов объектных моделей. Разграничены два типа моделей: порождающие и преобразующие. Уточняется критерий полноты для больших баз данных. Введены два критерия внутреннего совершенства моделей: декларативная и процедурная простота. Предложены количественные меры качества для порождающих и преобразующих моделей.

***Ключевые слова: объектные модели, метамодели, декларативная простота, процедурная простота, расстояние между текстами.***

**Kravchuk I. S. Linguistic Metamodels.** The paper is devoted to some estimation methods of alternative linguistic models of different types. A system of criteria of such estimation offering a choice of a certain model among the competing ones is named as a metamodel. The paper, depending on variety sets of objective models, studies some different types of the metamodels. It also shows how to specify a criterion of completeness in the large databases; distinguishes between two types of the models – generating and transforming ones; brings in the both criteria of the models' internal perfection – declarative and procedural simplicities; and provides quantitative measures of quality for the generating and transforming models.

***Key words: objective models, metamodels, declarative simplicity, procedural simplicity, distance between texts.***

У комп'ютерній лінгвістиці для розв'язання будь-якої задачі може бути запропоновано кілька різних моделей (алгоритмів). При цьому виникає проблема критеріїв оцінки альтернативних моделей і вибору оптимальної з деякої точки зору. Систему таких критеріїв оцінки і вибору однієї з кількох конкуруючих моделей можна назвати метамоделлю (пор. з об'єктною теорією і метатеорією).

Розгляньмо різні типи метамоделей в залежності від різних видів об'єктних моделей.

Першу спробу сформулювати вимоги до «ідеальної» лінгвістичної теорії зробив засновник глосематики Л. Єльмслев. Частина цих вимог характеризує відношення оцінюваної об'єктної моделі до реальності. Вони характеризують зовнішню виправданість моделі. Інша частина вимог стосується внутрішньої структури оцінюваної моделі. Вони характеризують внутрішню довершеність побудованої моделі.

Зовнішня виправданість включає: 1) вимогу забезпечити розуміння свого об'єкту [3:274] і 2) придатність лінгвістичної теорії. Перша вимога в генеративній граматиці називається пояснювальною силою, а в теорії моделей –екстраполяцією моделі, тобто здатністю прогнозувати не тільки мовний матеріал, використаний при побудові моделі, але й потенційно можливий матеріал, прийнятний для носіїв даної мови. Прийнятність означає відповідність експериментальним даним, використовуваним при верифікації моделі. Ідеальна модель має прогнозувати всі «правильні» лінгвістичні об'єкти і не прогнозувати «неправильні». Таким чином, стає очевидною необхідність розщепити поняття придатності моделі на два поняття: повноти й адекватності.

Внутрішня довершеність побудованої об'єктної моделі, крім вимоги несуперечності, включає також вичерпний характер моделювання й простоту. Що стосується вичерпного характеру, то по суті ця ознак збігається з повнотою. Що ж стосується простоти, то слід розрізняти її види. Так, Г. Рейхенбах розрізняв дескриптивну й індуктивну простоту [8:140–141].

Дескриптивна простота – це простота усередині еквівалентних описів, а індуктивна – це здатність моделі за допомогою невеликої кількості понять описувати величезну кількість спостережуваних фактів [8: там же]. Таке розмежування є важливим з точки зору логіки науки, але недостатнім з точки зору кількісної оцінки якості об'єктної моделі, що саме й становить мету побудови метамоделей. Для оцінки якості моделювання слід враховувати два типи знань, використовуваних, взагалі кажучи, в будь-якій моделі: декларативні й процедурні. Вони відповідають двом способам задання множин: переліком і описом. Таким чином, декларативні знання – це спискове задання фактів переліком, а процедурні – це задання їх у вигляді процедур, за допомогою яких їх можна отримати. Відповідно до цього і поняття простоти можна розщепити на два поняття: декларативну і процедурну простоту.

Деякі зі згаданих понять були формалізовані І. О. Мельчуком [6:113–123]. Вони стосуються одного з типів лінгвістичного опису – породжуючих граматик. Крім того, вони виходять з ідеальної ситуації, коли дослідник має вичерпну інформацію, по-перше, про всі елементи, породжувані моделлю, і, по-друге, про всі елементи, призначені для породження. Така ситуація рідко трапляється на практиці. Тому згадані поняття придаються «… не столько для того, чтобы вооружить читателя средством проверки качества лингвистических теорий, сколько для того, чтобы продемонстрировать принципиальную возможность строго говорить о категориях, которые на первый взгляд не поддаются формализации» [1:268].

Наведені слова можна віднести і до широко розповсюджених методів кількісної оцінки технічної ефективності пошуку в повнотекстових базах даних. Дж. Перрі, А. Кент і М. Беррі запропонували як параметр функціонування ІПС коефіцієнт повноти (recall factor) [9; 10], визначуваного як відношення виділених релевантних документів до загальної кількості всіх релевантних документів у масиві. Але встановити цю загальну кількість можна лише в масивах невеликого обсягу і практично неможливо для фондів Інтернету. У зв'язку з цим треба внести зміни у використовувані формули оцінки якості лінгвістичних моделей.

Нехай число результівних лінгвістичних елементів, видаваних даною моделлю (наприклад, породжуючою або інформаційно-пошуковою), дорівнює N, а число елементів, визнаних експериментатором релевантними, дорівнює R. Тоді повноту *C* (completness) функціонування моделі можна подати як відношення числа релевантних елементів до загальної кількості всіх виданих елементів (а не всіх потенційно можливих елементів!), виражене у відсотках:

$$C = 100*R / N \qquad (1).$$

Точність *P* (precision) оцінюваної моделі у свою чергу можна подати як відношення числа $(N – R)$ нерелевантних елементів до загальної кількості всіх виданих елементів, виражене у відсотках:

$$P = 100*(N – R) / N \qquad (2).$$

Звернімо увагу, що формули (1) і (2) застосовні також для оцінки якості функціонування породжуючи моделей, алгоритмів аналізу й синтезу текстів, автоматичних ІПС, включаючи й так звані «пошукові машини» Інтернету.

Якщо в нашому розпорядженні є відкритий початковий код програми, яка реалізує дану модель, то її можна оцінити з точки зору внутрішньої довершеності, інакше кажучи, з точки зору декларативної й процедурної простоти.

Позначимо через *D* число даних програми, оголошених декларативно, а через *O* – число операторів програми. Тоді декларативну простоту *S(D)* моделі можна виміряти формулою:

$$S(D) = C / D \qquad (3),$$

а процедурну простоту *S(O)* – формулою

$$S(O) = C / O \qquad (4).$$

Максимальне значення *S(D)* і *S(O)* дорівнює 100. Таке значення ці величини набувають, якщо в програмі використовується лише одна змінна і один оператор (наприклад, друк змінної, як у деяких навчальних програмах), і при цьому програма видає лише один релевантний результат. Мінімальне значення *S(D)* і *S(O)* дорівнює 0, якщо програма не видає жодного релевантного результату і повнота видачі *C* дорівнює 0.

Як свідчить практика, дуже часто між повнотою *C* і точністю *P*, а також між декларативною простотою *S(D)* і процедурною *S(O)* існує зворотна залежність. Так, наприклад, при побудові алгоритму синтезу відмінкових форм складених числівників української мови набагато доцільніше виходити зі списку готових форм компонентів цих числівників, ніж будувати компоненти зі списку відповідних морфем. При цьому за рахунок неістотного збільшення декларативної інформації істотно спрощується процедурна частина алгоритму.

Аналогічні методи побудови метамоделей використовуються для оцінки якості автоматичного реферування. Наразі здійснено доволі багато спроб розроблення формальних критеріїв для такої оцінки [5]. Разом з тим існує можливість установлення деяких загальних принципів побудови метамоделей автоматичного реферування, таких, як:

1) установлення формального ознакового простору для характеристики потрібних властивостей рефератів;

2) установлення еталонних значень запропонованих ознак;

3) установлення властивостей даного реферату шляхом порівняння значень його ознак з еталонними значеннями;

4) сумарна оцінка якості реферату на підставі комплексного показника якості.

Найважливішим етапом, як у всіх задачах розпізнавання образів і кластеризації, є перша задача. Ознаки реферату можна розділити на: 1) мовні, 2) структурні й 3) оформлювальний. Наприклад, до мовних можна віднести: насиченість тексту термінами, відсутність неінформативної лексики, наявність показників семантичних відношень, наявність простих речень у тексті. До структурних ознак слід включити дотримання регламентованої структури побудови тексту. До оформлювальних ознак можна включити наявність шрифтового оформлення та графічних засобів.

Як еталон значення ознаки умовно можна обрати одне з крайніх значень показника. Наприклад, для показника насиченості термінами як еталон обирається стовідсоткова насиченість, тобто ситуація, при якій усі поняття реферату представлені термінами з інформаційно-пошукового тезаурусу. Для показника надлишковості як еталон можна використати повну відсутність у тексті надлишкової лексики (стоп-слів). Шляхом порівняння абсолютного значення ознаки реферату з еталонним значенням встановлюється відносне значення відповідної ознаки за формулою:

$$V_{[r]} = V_{[i]} / V_{[et]} \qquad (5),$$

де $V_{[r]}$ – відносне значення певної ознаки, $V_{[i]}$ – абсолютне значення *i*-тої ознаки, $V_{[et]}$ – еталонне значення ознаки.

Комплексний показник якості реферату у найпростішому випадку можна визначити як суму відносних значень його елементарних ознак.

За наведеною методикою було проведено порівняльний аналіз [7] трьох систем автоматичного реферування англомовних текстів: 1) «Автореферат» Microsoft Word 2003; 2) Intelexer Summarizer 2.6; 3) TextAnalist 2.01. Найнижчу оцінку якості 0.1249 отримала перша система; найвищу 0.2656 – третя система. Такі показники повністю узгоджуються з інтуїтивно-експертною оцінкою цих систем, що свідчить про достатню адекватність розглянутих об'єктивно-формальних критеріїв оцінки якості реферування.

Способи оцінок лінгвістичних моделей залежать, з одного боку, від зовнішніх і внутрішніх критеріїв оцінки, а з другого боку, від типу вхідних і вихідних даних моделі. В залежності від типу вихідних даних моделі можна розділити на: 1) моделі породження і 2) моделі перетворення. У перших вихідні дані можуть містити більше одного об'єкта (наприклад, множину всіх породжених словоформ, речень або множину документів, виданих по одному запиту), а у других – вихідні дані містять лише один об'єкт (наприклад, при перекладі кожному реченню зіставляється у тексті перекладу лише одне речення). Тому для оцінки якості машинного перекладу не можна застосувати розглянуті вище способи оцінки.

Оцінка якості МП може ґрунтуватися як на внутрішньому, так і на зовнішньому підході. У першому випадку оцінка включає такі властивості, як якість грубого (невідредагованого) перекладу, наприклад, зрозумілість, точність, стилістична адекватність; зручність засобів створення й поповнення словників; можливість використання для інших пар мов; обсяг постредагування й тощо [11:13.3].

Зовнішній підхід ґрунтується на порівнянні невідредагованого перекладу, виданого комп'ютером, і еталонного перекладу, здійсненого людиною. У результаті такого порівняння створюється параметричний опис об'єкта – МП-тексту. Як ознаки такого опису використовуються різні міри близькості (відстані) [4] між порівнюваними реченнями. Відстань між реченнями, взагалі беручи, можна визначати подвійно: процедурно й декларативно.

При процедуральному підході для кожного речення визначається число операцій, необхідних для того, щоб перетворити МП-речення в еталонне речення за допомогою припустимих операцій вставки, усунення і заміни однієї словоформи. При декларативному підході використовується число випадків збігу ***n***-грамних послідовностей словоформ [5] у порівнюваних реченнях. Сумарна оцінка відстаней між текстами обчислюється як середня величина відстаней між реченнями.

Лінгвістичні метамоделі грають важливу роль при створенні конкурентного середовища для розвитку комп'ютерної лінгвістики і, таким чином, є важливим засобом сприяння прогресу досліджень у даній галузі.

## Література

1. Апресян Ю. Д. Идеи и методы современной структурной лингвистики / Ю. Д. Апресян. — М. : Просвещение, 1966. — 301 с.

2. Горькова В. И. Реферат в системе научной коммуникации. Направления совершенствования лингвистических и структурных характеристик / В. И. Горькова, Э. А. Борохов // Итоги науки и техники. — М., 1989. — Сер. «Информатика». — Т. 11. — 232 с.

3. Ельмслев Л. Пролегомены к теории языка / Луи Ельмслев // Новое в лингвистике. — М. : ИЛ. — Вып. 1. — 1960. — С. 264—280.

4. Кравчук И. С. О функциях принадлежности заданному типу текстов / И. С. Кравчук // Вісник ХНУ ім. В. Н Карзіна. : Сер. Філологія. — 2002. — № 572, вип. 36. — С. 24—29.

5. Мани Индерджиет. Система автоматического реферирования [Электронный ресурс] / Индерджиет Мани, Удо Хан. — Режим доступа : http://www.osp.ru/os/2000/12/178370/_p4.html

6. Мельчук И. А. О стандартной форме и количественных характеристиках некоторых лингвистических описаний / И. А. Мельчук // Вопросы языкознания. — 1963. — № 6. — С. 113—123.

7. Носач Д. С. Методи оцінки ефективності систем автоматичного реферування : [дипломна робота] / [кер. доц. Кравчук І С., доц. Кириленко О. Г. ] / Д. С. Носач. — Харків : Аерокосмічний університет імені М. Є. Жуковського «ХАІ», 2009. — 47 с.

8. Шаумян С. К. Структурная лингвистика / С. К. Шаумян. — М. : Наука, 1965. — 395 с.

9. Cleverdon C. W. The testing of index language devices / C. W. Cleverdon and J. Mills // Aslib Proceedings. — 1963. —v. 15. —№ 4. — 106—130 p.p.

10. Perry J. Machine literature searching / J. Perry, A. Kent and M. Berry. — New York, Interscience Publishers, Inc., 1956. — 42—49.

11. Survey of the State of the Art in Human Language Technology / Center for Spoken Language Understanding. — Oregon Graduate Institute, USA; University of Pisa, Italy. — Режим доступу : http://www.cslu.ogi.edu/HLTsurvey/HLTsurvey.html.