РОЗДІЛ IV. МОДЕЛЮВАННЯ ЛЕКСИЧНИХ І ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ

Данилюк Ілля

УДК 81’33

КОЛЬОРОВА МАПА РОМАНУ «КРИНИЧАР» МИРОСЛАВА ДОЧИНЦЯ У *MATHEMATICA*

У статті описано ідею та реалізацію процесу створення кольорової мапи (КМ) тексту на основі тексту роману «Криничар» Мирослава Дочинця. КМ – це множина кольорових квадратиків (або інших фігур), кожен із яких представляє конкретну кольороназву в оригінальному тексті. Цілком об’єктивний результат унаочнює розподіл у тексті конкретних прикметників на позначення кольору. Оригінал статті написано у форматі Computable Document Format (CDF) – обчислюваний документ, і може бути використаний для довільного тексту українською мовою, з урахуванням особливостей відмінювання і навіть словотвору.

Ключові слова: кольороназва, мовна модель, словозміна.

Сьогодні є чимало інструментів для опрацювання природного мовлення, зокрема, з високим ступенем автоматизації. Сучасні мови програмування дозволяють ґрунтовно обробляти текстові дані на рівні окремих символів, групи символів (фраз і речень), у перспективі мають з'явитися функції опрацювання семантики. Кількість вбудованих рядкових функцій зростає, і це розширює потенціал розробника і допомагає заощаджувати час, необхідний для власноручного створення таких функцій і процедур. Наприклад, процедуру визначення різниці між двома рядками – так званої мінімальної відстані редагування або числа Левенштейна – розробник на Python має написати вручну [Jurafsky 2009]. Втім, у деяких сучасних систем ця процедура входить до вже вбудованих. Обчислювальне програмне середовище *Mathematica* від Wolfram Research, яке широко використовується в багатьох наукових, інженерних, математичних і обчислювальних проектах [Wellin 2013], а нами було застосоване у галузі лінгвістики, наприклад, має вбудовану функцію *EditDistance [X, Y]*. На нашу думку, основні алгоритми автоматичного опрацювання природного мовлення, як-от лематизація, синтез словоформ, визначення граматичних класів і підкласів, синтаксичного аналізу або навіть автоматизованого перекладу, які описано в [Баранов 2003; Волошин 2004; Дарчук 2008; Карпіловська 2006; Карпіловська 2006; Партико 2008], згодом стануть вбудованими процедури.

Ця стаття в оригіналі написана у форматі CDF – *обчислюваний документ* для системи *Mathematica*, або *Notebook* (*.nb) – і її можна завантажений [Данилюк 2014a], оскільки код у друкованій версії буде наведено тільки частково. Для перегляду CDF вам знадобиться безкоштовна програма від [http://www.wolfram.com/cdf-player/].

Отже, головна мета статті полягає в описі процесу, інструментів і безпосередньо коду для автоматичного генерування кольорової мапи для довільного тексту українською мовою, і зокрема, для тексту роману «Криничар» Мирослава Дочинця. Кольорова мапа (КМ) – це множина кольорових квадратиків (або інших фігур), кожен із яких представляє конкретну кольороназву в оригінальному тексті. Кожне використання прикметника на позначення кольору – «білий», «чорний», «червоний», «золотий» – у КМ буде представлено квадратиком відповідного кольору. Отже, можна отримати повне й абсолютно об'єктивне представлення лексики конкретного тексту і певні риси «картини світу», концепти окремих кольорів у літературних творах. Це питання є досить актуальним в українській лінгвістиці і філології, тому інструмент для автоматизованого пошуку кольороназв у довільному текстів, на нашу думку, є необхідним.

КМ можна вважати автоматизованою інфографікою для візуалізації мовленнєвих даних. Загальна ідея належить Тетяні Дружняєвій із видання «Esquire», окремі елементи коду запропоновано Романом Осиповим (Московський державний університет тонких хімічних технологій).

Ми поділили дослідження на **декілька завдань**: 1) отримати всю можливу статистичну інформацію з тексту для подальшого аналізу; 2) створити процедури (мовою для *Mathematica*), щоб працювати з окремим словами та реченнями; 3) побудувати мовну модель для називання кольору з урахуванням відмінювання й окремих випадків словотвору; 4) використати модель для генерування КМ тексту роману «Криничар» Мирослава Дочинця, і описати перспективи.

Об'єкт дослідження – текст роману «Криничар» Мирослава Дочинця. Формально це текстовий файл формату txt, підготовлений для обробки (в Unicode, кожен токен відділений пробілом). Предметом дослідження є використання кольороназв, представлених у вигляді КМ.

Розпочнемо з першим завданням. Текстовий файл – stus.txt – має бути в тій самій папці, що й робочий файл *.nb, і ми читаємо дані з нього у змінну stus:

```
Short[stus=Import[FileNameJoin[{NotebookDirectory[],"stus.txt"}]]]
```

Коли ми звертаємося до stus, то працюємо з усім текстом і можемо отримати основні статистичні дані – скільки символів він містить:

```
StringLength[stus]
579771
```

або скільки рядків:

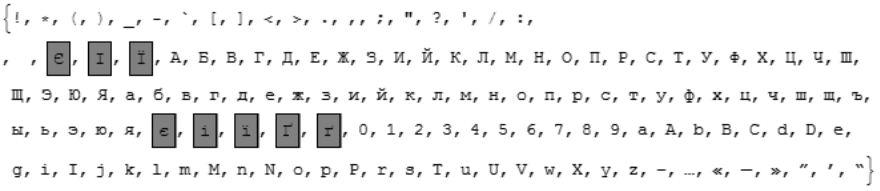
```
StringCount[stus,"\n"]
23293
```

або скільки приблизно речень (за умови, що речення може закінчується крапкою, знаком оклику або знаком питання):

```
StringCount[stus,{".","?","!"}]
10689
```

або які символи використовуються в тексті, у відсортованому вигляді:

```
Union[Characters[stus]]
```



Зверніть увагу на те місце, де опинилися після сортування деякі специфічні українські літери (їх виділено). Причина в тому, що в стандартній таблиці символів вони розташовані не у загальному списку кирилиці, а на випадкових позиціях. Цю незручністьнеобхідно враховувати у разі використання регулярних виразів (РВ) на кшталт "всі українські літери від А до Я". На практиці РВ /[Є-ґ]+/ досить добре для підходить для запиту "всі українські літери", але він повертає кілька не-українських символів – *Э, ъ, ы, э*.

Наступним кроком переходимо до обробки слів. Змінна *allWords* містить всі словоформи зі змінної *stus* без розрізнення великих і малих літер – через заміну за допомогою РВ:

```
Short[allWords = Sort[Tally[DeleteCases[StringSplit[StringReplace[StringReplace[stus,
Thread[Join[CharacterRange["А", "Я"], {"Є", "І", "Ї", "Ґ"}] -> Join[CharacterRange["а", "я"], {"є", "і", "ї", "ґ"}]]],
RegularExpression["[^" <> StringJoin@Join[CharacterRange["а", "я"], {"є", "і", "ї", "ґ"}] <> "]"] -> " "], " "], ""]],
#1[[2]] > #2[[2]] &], 5]
```

Кількість токенів у *stus* сягає 93442, а унікальних словоформ – 23182. Ось 100 найчастотніших:

```
allWords[[1";;" 200]][[";;",1]]
{"і", "я", "не", "на", "в", "з", "а", "що", "й", "як", "до", "у", \
"за", "це", "його", "він", "мене", "так", "мені", "коли", "від", \
"ти", "їх", "для", "то", "ми", "бо", "але", "ще", "та", "під", "аби", \
"її", "вони", "із", "тоді", "про", "чи", "по", "вона", "тут", "було", \
"був", "ні", "вже", "лише", "собі", "той", "себе", "очі", "все", \
"тебе", "них", "там", "гроші", "ж", "де", "тобі", "сам", "того", \
"може", "йому", "якщо", "щось", "нього", "мій", "те", "б", "хто", \
"люди", "цього", "тепер", "навіть", "нас", "більше", "руки", "є", \
"щоб", "чоловік", "через", "треба", "мав", "тому", "свою", "буде", \
"були", "над", "ви", "свої", "їм", "при", "добре", "аж", "далі", \
"час", "вода", "людей", "пан", "воду", "без"}
```

Для створення можливості шукати конкретну словоформу в тексті будуємо функцію *wordPosition*, яка повертає масив символів позицій.

```
replacements=Thread[Join[CharacterRange["а","я"],{"є","і","ї","ґ"}]-
>Join[CharacterRange["А","Я"],{"Є","І","Ї","Ґ"}]]
```

Ось результат для слова *тополя*:

```
wordPosition["тополя"]
{{219414, 219421}, {408610, 408617}, {426728, 426735}, {484326, 484333}}
```

Якщо ми знайдемо позиції для крапок (та інших символів у кінці речення – – “!”, “?”, “…”), можна буде отримати ціле речення (послідовність символів між двома розділовими знаками) і записати у змінної *sentence*.

```
Short[dots = #[[1]] & /@ StringPosition[stus, {".","?","!","\[Ellipsis]"}], 5];
sentense[{min_,max_}]:=Block[{start=Select[Nearest[dots,min,10],#<min&][[1]]+1,end=Select[Nearest[dots,m
ax,10],#>max&][[1]]},StringTake[stus,{start,end}]]
```

Поєднання *wordPosition* і *sentence* виведе конкорданс для конкретної словоформи:

```
Grid[Transpose@{StringReplace[sentense /@ wordPosition["тополя"],
"\n" -> ""]}, Background -> {None, {{Orange, LightGray}}},
ItemStyle -> Directive[16, Bold, FontFamily -> "Arial"],
Alignment -> Left, Dividers -> All]
```

Тепер треба визначити загальну колірну модель для побудови КМ. Її релевантність і глибина для української мови – зокрема, тічність лематизації і докладність – мають вирішальне значення для якості КМ. Перший крок – знайти прикметники, які прямо позначають кольори і складаються з однієї словоформи. Заносимо ці прикметники у масив *tc* (таблиця кольорів) і описуємо їх за моделлю RGB (фрагмент коду досить

довгий і підходить тільки для перегляду в цифровій версії). Це прикметники: *білий, червоний, зелений, синій, жовтий, чорний, сірий, рожевий, коричневий, блакитний, пурпурний, пурпуровий, оранжевий, помаранчевий, фіолетовий, амарантовий, бурштиновий, аметистовий, абрикосовий, аквамариновий, арсеновий, спаржевий, бежевий, латунний, бронзовий, брунатний, карміновий, морквяний, лазуровий, каштановий, шоколадний, цинамоновий, кобальтовий, мідний, кораловий, кукурудзяний, блаватний, кремовий, малиновий, джинсовий, смарагдовий, баклажановий, лляний, золотий, індиго, нефритовий, хакі, лавандний, лимонний, бузковий, малахітовий, гірчичний, оливковий, помаранчевий, ліловий, персиковий, грушевий, барвінковий, сливовий, бурий, іржавий, шафрановий, сапфіровий, багряний, срібний, болотний, мандариновий, будяковий, бірюзовий, ультрамариновий, фіолетовий, пшеничний.*

А ось фрагмент представлення *tc*:

білий	GrayLevel[1]
червоний	RGBColor[1, 0, 0]
зелений	RGBColor[0, 1, 0]
синій	RGBColor[0, 0, 1]
жовтий	RGBColor[1, 1, 0]
чорний	
сірий	GrayLevel[0.5]
рожевий	RGBColor[1, 0.5, 0.5]
коричневий	RGBColor[0.6, 0.4, 0.2]
блакитний	RGBColor[0, 1, 1]
пурпурний, пурпуровий	RGBColor[1, 0, 1]
оранжевий, помаранчевий	RGBColor[1, 0.5, 0]
фіолетовий	RGBColor[0.5, 0, 0.5]
амарантовий	RGBColor[229/255, 43/255, 16/51]

Колірна модель також включає в себе правила для побудови різних словоформ і пошуку їхніх лем. Так, *червоний, червоного, червона, червонії* тощо будуть представлені лемою *червоний* і червоним квадратиком. Для цього використовуємо синтез словоформ за словником основ і закінчень (отримання всіх можливих словоформ для кожного прикметника в таблиці кольорів), потім знаходимо позиції для цих словоформ у змінній *stus*, присвоюємо ці позиції конкретній лемі. Нарешті, діапазон лем, як вони з'являються в тексті, замінюємо на кольорові квадратики.

Відмінювання прикметника в українській мові включає тверду (*червоний*) і м'яку групи (*синій*), стягнені (*червона*) і нестягнені форми (*червоная*). Змінні *ColorEdningsTv* і *ColorEdningsMk* містять відповідно закінчення для стягнених і нестягнених форм твердої групи та стягнених і нестягнених форм м'якої групи:

ColorEdningsTv={"ий","ого","ому","им","ім","е","а","ої","ій","у","ою","і","их","им", "ими", "ає","ая","ую","ії"};

ColorEdningsMk={"ій","ього","ьому","ім","є","я","ьої","ю","ьою","і","іх","іми","єє", "єє","яя", "ою","ії"};

Змінна *colorRules* містить згенерований масив усіх можливих словоформ для називання кольорів (поєднання основ із *tc* і закінчень із *ColorEdningsTv* і *ColorEdningsMk*) зі вбудованим механізмом врахування афіксів (*білий – біленький*). Фрагмент коду:

```
colorRules=Flatten[{Thread[Flatten[Table[{"біл",
"біленьк"}[[v]]<>ColorEdningsTv[[u]],{v,2},{u,Length[ColorEdningsTv]}]]->White],
Thread[Table["син"<>ColorEdningsMk[[u]],{u,Length[ColorEdningsMk]}]->Blue],…
Thread[Table["пшеничн"<>ColorEdningsTv[[u]],{u,Length[ColorEdningsTv]}]-
>RGBColor[245/255,222/255,179/255]]}]
```

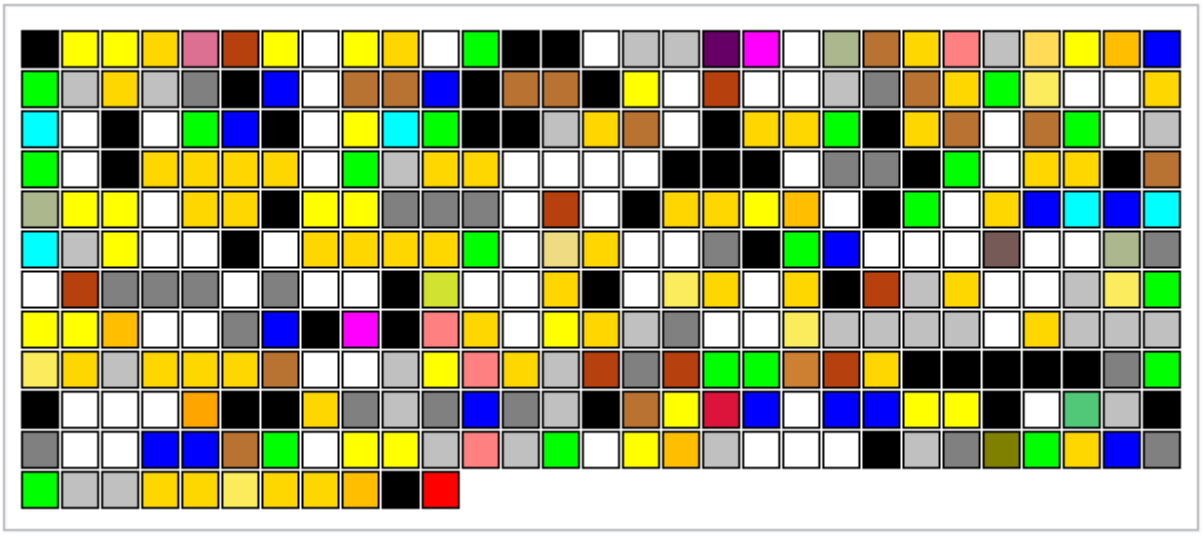
Позиції згенерованих словоформ записуємо у змінній *colorInformationPre*, а потім сортуємо у *colorInformation*:

```
colorInformationPre={#,wordPosition[#]}&/@colorRules[[;;,1]];
colorInformation=Sort[Flatten[Partition[Riffle[#[[2]],#[[1]],{2,-1,2}],2]&/
@colorInformationPre,1],Mean[#1[[1]]]<Mean[#2[[1]]]&]
```

І, нарешті, будуємо КМ:

```
Panel[Grid[Partition[Graphics[{#,EdgeForm[Black],Rectangle[]},ImageSize-
>20]&/@(colorInformation[[;;,2]]/.colorRules),30,30,1,""],Spacings->{0,0}]]
```

Ось результат:



Текст роману є порівняно багатим на кольороназви (Данилюк 2014б; Danyliuk, Данилюк 2016), основними з яких є жовтий/золотий, чорний і білий, відтінки сірого. Практично немає червоного, незначна кількість зеленого, синього і блакитного. Порівняння КМ «Криничара» з КМ інших творів літератури і фольклору може бути темою для подальшого вивчення. Як пише автор дослідження [Ковтун 2009: 51], «колірна ознака з'явилася в мові в діахронічній послідовності. У народній творчості спочатку переважають означення білого та чорного кольорів, за ними йде червоний (тріада білий — чорний — червоний), після нього — зелений і жовтий, далі — синій і брунатний». Виглядає, що «Криничар» Мирослава Дочинця в аспекті використання кольороназв є оригінально-авторським.

Описаний спосіб побудови КМ на основі позиції лексеми на позначення кольору є базовим. Його можна поліпшити через додавання до колірної моделі похідних прикметників (*білявий, чорнющий*), дієслів (*біліти*) та іменників (*чорнота*), деяких назв вторинних відтінків (*ясно-зелений*), тісно пов'язаної лексики – *вороний, гнідий* тощо.

Роман, крім того, представлений у вигляді корпусу текстів на ресурсі corpora.pp.ua.

Література

Баранов 2003: Баранов А. Н. Введение в прикладную лингвистику [Текст] / А. Н. Баранов. – М. : Едиториал УРСС, 2003. – 360с. – ISBN: 5-8360-0196-0.

Волошин 2004: Волошин В. Г. Комп'ютерна лінгвістика : Навчальний посібник [Текст] / В. Г. Волошин. – Суми : ВТД «Університетська книга», 2004. – 382с. – ISBN: 966-680-134-5.

Данилюк 2014а: Данилюк І. Аналіз тексту "Кобзаря" Тараса Шевченка в середовищі Mathematica: символи, слова і кольори [Text]. – Access mode : URL : https://app.box.com/stus. – Title from the screen.

Данилюк 2014б: Данилюк І. Кольорова мапа поетичного спадку Василя Стуса у mathematica [Текст] / І. Данилюк // Вісник Донецького національного університету: Серія Б. Гуманітарні науки. — 2014. — Т. 1-2. — С. 78–84.

Данилюк 2016: Данилюк І. Кольорова мапа трилогії «Волинь» Уласа Самчука у mathematica [Текст] / І. Данилюк // Лінгвокомп'ютерні дослідження. — 2016. — №. 9. — С. 111–121.

Дарчук 2008: Дарчук Н. П. Комп'ютерна лінвістика: Автоматичне опрацювання тексту [Текст] / Н. П. Дарчук. – К. : Видавничо-поліграфічний центр «Київський університет», 2008. – 351с. – ISBN 978-966-439-079-5.

Карпіловська 2006: Карпіловська Є. А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика. Підручник [Текст] / Є. А. Карпіловська. – Донецьк : ТОВ «Юго-Восток, Лтд», 2006. – 188с. – ISBN 966-374-078-7.

Ковтун 2009: Ковтун Л. Український колористичний код світотворення // Вісник Київського національного університету імені Тараса Шевченка. Українознавство. – Вип. 13 / КНУ імені Тараса Шевченка. – Київ : ВПЦ "Київський університет", 2009. – ISSN 1728-2330.

Марчук 2000: Марчук Ю. Н. Основы компьютерной лингвистики [Текст] / Ю. Н. Марчук. – М. : Народный учитель, 2000. – 320с. – ISBN 5-17-039480-2.

Партико 2008: Партико З. В. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності [Текст] / З. В. Партико. – Львів: Афіша, 2008. – 224 с. – ISBN 978-966-325-092-2.

Danyliuk: Danyliuk I. Color Map of Taras Shevchenko's «Kobzar» with Mathematica / I. Danyliuk [Текст] // Linguistic Studies. — 2014. — Т. 29. — С. 218–223.

Jurafsky 2009: Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Текст]. – Prentice Hall, 2009. – 988p. – ISBN 0-13-095069-6.

Wellin 2013: Wellin Paul R. Programming with Mathematica, An Introduction. – Cambridge, 2013. – 711 pp. – ISBN: 9781107009462.

В статье описана идея и реализация процесса создания цветовой карты текста на основе текста романа «Криничар» Мирослава Дочинца. Цветовая карта — это множество цветных квадратиков (или других фигур), каждый из которых представляет конкретное слово для обозначения цвета в оригинальном тексте. Весьма объективный результат наглядно представляет распределение в тексте конкретных прилагательных для обозначения цвета. Оригинал статьи написан в формате Computable Document Format (CDF) — вычисляемый документ, и может быть использован для произвольного текста на украинском языке, с учетом особенностей склонения и даже словообразования.

Ключевые слова: название цвета, языковая модель, словоизменение.

The article describes an idea and its realization process for creating Color Map (CM) for the text – in particular "Krynychar" by Myroslav Dochynets. CM is a composition, a grid made of colored rectangles (or any other figures) – one for every name of color in the original text. Absolutely objective result shows visually the distribution of particular adjectives. The original article is in Computable Document Format (CDF) and is suitable for random text in Ukrainian considering inflection and even derivation.

Keywords: color name, language model, inflection.

Надійшла до редакції 18 січня 2016 року

Жанна Краснобаєва-Чорна

УДК 811.162.2'373.7

СПЕЦИФІКА МОДЕЛЮВАННЯ НЕГАТИВНООЦІННИХ ФРАЗЕМ (НА МАТЕРІАЛІ УКРАЇНСЬКОЇ, РОСІЙСЬКОЇ Й АНГЛІЙСЬКОЇ МОВ)

Стаття продовжує цикл публікацій автора, присвячених проблемам фраземної аксіології. Аналіз моделювання негативнооцінних фразем у статті має описовий характер. Зафіксовано наявність фразем із типовою синтаксичною структурою й однаковою семантикою, що уможливлює виділення структурно-семантичних моделей негативноо цінних фразем. Специфіка фраземного моделювання семантичних діапазонів «погано», «не схвалювати», «не «задовольняти», «не цікавити», «заперечувати», «не відповідати нормі» відзначається нерегулярністю.

Ключові слова: аксіологія, оцінка, фразема, негативнооцінна фразема, фраземна модель.

У вітчизняній і зарубіжній лінгвістиці активно досліджуються теоретичні та практичні питання фраземного моделювання (див. праці Ю. Божко [Божко 2002], Г. Бабаєвої [Бабаева 2011], В. Губарева [Губарев 1985], Д. Давлетбаєвої [Давлетбаева 2012], Ж. Краснобаєвої-Чорної [Краснобаєва-Чорна 2015б], Л. Молчанової [Молчанова 2013], Г. Ситар [Ситар 2012], Г. Солганика [Солганик 1976] та ін.). Проте не досить опрацьованим видається аксіологічний аспект моделювальної діяльності у фраземіці, чим і зумовлена актуальність статті.

Мета статті полягає у виявленні особливостей процесу моделювання негативнооцінних фразем. Джерельну базу дослідження становлять фраземи, почерпнуті з авторитетних фраземографічних видань української (СФУМ 2003), російської (ФСРЯ 1987), англійської (Кунин 2000) мов.

Негативнооцінні фраземи позиціоновано в статті як фраземи, семантика яких відповідає зоні «погано» оцінної шкали й інтерпретована як «погано», «не схвалювати», «не «задовольняти», «не цікавити», «заперечувати», «не відповідати нормі» [Краснобаєва-Чорна 2015а]. Фраземну модель дефіновано як структурно-семантичний інваріант стійких сполучень, що семантично відображає відносну стабільність їхньої форми та семантики [Мокиенко 1989: 53]. У мовознавстві можна виділити три погляди на проблему моделювання у фраземіці:

1) думки про немоделювальну сутність фразем, тобто заперечення можливості виникнення фразем за певною моделлю з прогнозованим значенням, дотримуються Н. Амосова, В. Телія, І. Чернишева та ін.;

2) моделювальний характер фразем визнають Ю. Бурмістрович, К. Богатирева, С. Гаврин, В. Губарев, Ю. Гвоздарьов і ін. Так, на думку Ю. Бурмістровича, утворення фразем здійснюється з використанням знань про відповідності між типами значень і типами структур фразем, а також між типами семантичних відносин між фраземами та їхніми «виробниками» [Бурмистрович 1971: 12].

У межах цього погляду актуалізовано питання про характер фраземного моделювання – породжувальний або описовий. Під час розгляду фраземних моделей, на думку К. Богатиревої, слід враховувати граматичний лад і семантику фраземи. Досліджуючи специфіку фраземного моделювання в синхронії та діахронії, лінгвіст наголошує на особливому характері структурно-семантичних моделей фразем: з одного боку, фраземні моделі в синхронії є моделями опису типологійних особливостей структури та семантики вже наявних фразем; з іншого боку, структурно-семантичні схеми виконують породжувальну функцію в аспектах діахронії та динаміки, що пояснює виникнення одноструктурних фразем, об'єднаних певними семами. Типізація властивостей структурно-семантичних моделей фразем не обмежується описом особливостей їхньої семантичної структури в статиці [Богатырева 2015: 61]. Динамічний характер фраземних моделей, що виявляється в схильності до