DOI: 10.46340/eppd.2020.7.4.17

Eduard Kotov
ORCID ID: https://orcid.org/0000-0001-8372-7616
Vasyl' Stus Donetsk National University, Ukraine

SOME COMMENTS ON THE ALGORITHM OF AUTOMATIC ANALYSIS OF POLITICAL DISCOURSE

Едуард Котов
Донецький національний університет імені Василя Стуса, Україна.

ДЕЯКІ ЗАУВАЖЕННЯ ЩОДО АЛГОРИТМУ АВТОМАТИЧНОГО АНАЛІЗУ ПОЛІТИЧНОГО ДИСКУРСУ

The article deals with the process of implementing a linguistic program analysis of political discourse, indicating the possibility of automatically building a network model. It is determined that political discourse (as well as any type of discourse) can be represented in the form of an interconnected system of concepts that can provide a fairly comprehensive idea of the organization of political space within a certain political entity. The author substantiates the possibility of automatic analysis of the linguistic corpus and highlights the most important gaps that occur in the way of formalization of this type of analysis. The main technical stages of automatic discourse research are defined. It is proved that formal analysis is possible when the theoretical aspects are brought into epistemological correspondence with the technical possibilities of implementation.

Keywords: political discourse, discourse analysis, formal analysis, post-structuralism, linguistic analysis.

Постановка проблеми. Початкова позиція, заснована на пост структуралістській теорії дискурсу передбачала наявність відносно закріпленої структури знаків, які організовують дискурсивний простір і визначають все те, що може бути сказано і те як це «все» може бути сказано, щодо тих чи інших тем (детальніше ці питання підіймаються у наших інших публікаціях).

Загальним чином, зазначимо, що дискурс, організований навколо ключових понять (структуруючих знаків, вузлових точок), які не мають фіксованого смислу, але є тим, що структурує смислове поле, привласнюючи ідентичність, «зупиняючи ковзання означуючих, що фіксують їх значення»[1].Такі точки, як уже було сказано, привласнюють певні зафіксовані значення інших елементів ланцюга, семантично обмежуючи всілякі інтерпретації тих чи інших внутрішньо дискурсивних елементів. Загальне припущення полягає в тому, що подібні зв'язки можна вивести практично з будь-якого тексту «якщо читати його досить уважно і довго»[2]. Втім, безумовним для нас (і не тільки для нас[3]) бачився той факт, що існує пряма залежність обсягу тексту до виразності дискурсивної структури, яку можна з нього отримати. Дискурсивна структура, як нам здавалося, може бути виокремлена з тексту і буде представляти вид ієрархічно організованої мережі, в якій одні знаки, є «головними» (в теоретичному плані) і виступають основою для подальшого розгортання структури, яка виражається і доповнюється «плаваючими означаючими», ідентичність яких відкрита і формується в зчленуванні з іншими «означаючими» в ланцюгах еквівалентності[4].

[1] Жижек, С. (2009). *Возвышенный объект идеологии*. Москва: Художественный журнал.
[2] Howarth, D., Torfing J. (editors) (2004). *Discourse theory in European politics: identity, policy, and governance*. Palgrave Macmillan: Basingstoke.
[3] Howarth, D., Torfing, J. (editors) (2004). *Discourse theory in European politics: identity, policy, and governance*. Palgrave Macmillan: Basingstoke.
[4] Жижек, С. (2009). *Возвышенный объект идеологии*. Москва: Художественный журнал.

Мета статті. Метою статті є визначення основних етапів формального аналізу політичного дискурсу та розгляд можливого алгоритму аналізу.

Виклад основного матеріалу. Зміст змінюваних етапів і робочих гіпотез алгоритму виділення структури політичного дискурсу, узагальнюючи, можна представити наступними зауваженням щодо виділення найбільш значущих концептів: це завдання планувалося вирішувати різними способами на різних етапах дослідження. Спочатку, нам уявлялося, що найбільш вагомі концепти будуть більш-менш яскраво виражені в частотній складовій. Втім, ми практично відразу відмовилися від даного підходу, зберігши лише те, що слова, які вжиті в тексті одного разу не приймаються до розгляду в якості структуруючого знака (вузлової точки). Також розглядалася теза щодо першої появи слова по відношенню до загальної кількості слів у тексті. У будь-якому випадку, їх наявність, як нам здавалося є досить явною і не прихованою латентно-семантичними внутрішньоструктурними зв'язками. А тому, метод їх виокремлення повинен був ґрунтуватися на виділенні найбільш великих категорій, які можна буде ранжувати залежно від можливого ступеня їх структурної значущості, що передбачає їх включеність і посередницьку роль у розумінні інших елементів (принаймні на рівні «контекстуального наповнення»). Відповідно до цих теоретичних передумов нам уявлялося, що в автоматичній формі, на підставі корпусу текстів ми зможемо виділити набір категорій і понять (який, втім, не буде занадто великий (основних передбачалося не більше десяти)), безумовно вимагаючий інтерпретації, але описуючий наявний політичний дискурс.

Процес реалізації даного підходу та розроблюваних алгоритмів проводився на матеріалах текстів звернень Президента України до Верховної Ради України з 2015 по 2018 рр. включно. Підсумком стали деякі як методологічні, так і методичні (технічні труднощі).

Технічна реалізація наявного алгоритму може бути представлена як проведення декількох етапів:

1. Перед підготовчий етап (на цьому етапі проводиться обробка масиву текстових даних із наданням йому прийнятної для аналізу форми);
2. Безпосередній аналіз (обробка текстових масивів із застосуванням існуючого алгоритму виділення мережевої структури політичного дискурсу);
3. Отримання вихідних даних та побудова моделі.

В рамках перед підготовчого етапу необхідно проводити роботу безпосередньо з текстом на природній мові, який має наступні рівні класифікації [1]:

– рівень речень (висловлювань) – синтаксичний рівень
– рівень слів (словоформ – слів у певній граматичній формі) – морфологічний рівень;
– рівень фонем (окремих звуків, за допомогою яких формуються і розрізняються слова) – фонологічний рівень (втім, фонологічний рівень для комп'ютерної обробки не настільки важливий, оскільки відповідає рівню символів, виражених в алфавіті).

Найбільш важливими для аналізу представляються перші два рівні. Проблема тут полягає в наступному: з одного боку, не існує принципової програмної складності для поділу тексту на синтаксичні і морфологічні одиниці; з іншого боку, в рамках великого обсягу текстових даних, при неможливості «ручного» аналізу текстових масивів з'являється проблема релевантного розкладання корпусу текстів на окремі синтаксичні одиниці.

Відносно ж морфологічного рівня, тут ключовою проблемою є словоформи, тобто, форми, які може приймати слово в тих чи інших текстуальних потребах. Сюди відноситься як проблема диференціювання по аспекту приналежності того чи іншого слова до певної частини мови, що дозволяє їй реалізовувати різні структурні відносини в рамках речення, так і відмінки слів, які при автоматичному аналізі можуть бути сприйняті як різні елементи, що, отже, може зчинити істотний вплив на результуючий аналіз.

Реалізація даних технічних моментів вимагає включення під етапів, тому, етап перед обробки включає в себе наступні елементи: токенізація, виділення стоп-словника (працюючого в різний час аналізу), лемматизація (або стемінг).

Токенізація являє собою розбиття тексту на аналізовані одиниці (токени) [2]. Одиницями можуть виступати абзаци, пропозиції, слова. Також цей етап включає в себе видалення пробілів, великих

[1] Большакова, Е., Воронцов, К., Ефремова, Н., Клышинский, Э., Лукашевич, Н., Сапин, А. (2017). *Автоматическая обработка текстов на естественном языке и анализ данных*. Москва: НИУ ВШЭ.

[2] Teodorescu, M. (2017). *Machine Learning Methods for Strategy Research. HBS Working Paper 18-011*. Harvard Business School.

літер, цифр і т.п. це первинний етап, який реалізується практично в будь-якому середовищі автоматичного статистичного аналізу і не представляє істотної складності.

Наступні етапи є менш однозначними, оскільки відносяться до лексичних інструментів і їх виконання може істотно вплинути на результати дослідження. Першим з таких є приведення всіх допустимих словоформ до єдиного значення. Оскільки в іншому випадку, автоматичний аналіз текстових масивів втрачає свою репрезентативність. Спробуємо уточнити, що ми маємо на увазі. Припустимо у нас є слово " інститут "(політичний). У різних відмінках, множині або однині термін буде приймати різні форми:» інституту«,» інститутам«,» інститутів " і т. п. Це не представляє проблеми для аналітика, який проводить аналіз вручну, оскільки всі слова будуть відзначені і віднесені до категорії "інститут", але при автоматичному аналізі можна зіткнутися з проблемою релевантного відображення даних, в яких кількістна складова слова буде значно знижена при статистичному підрахунку, оскільки на рівні аналізу програма розпізнає кожне зі слів як відмінне. Крім того, тут ми стикаємося з іншою проблемою полісемії (наявність у однієї одиниці мови декількох пов'язаних значень), синонімії (повний або частковий збіг значень різних одиниці), омонімії (збіг за формою двох різних за змістом одиниць) (докладніше див. тут [1]). Крім того суттєва проблема може бути в анафорі, яка полягає в заміщення слів, речень, фрагментів тощо іншими мовними одиницями [2]. Відносне вирішення даних проблем припустиме при використанні інструментів стемматизації і лемматизації. Стемінг дозволяє привести слова до єдиної форми шляхом відсікання (за заданими правилами) суфіксів від слова. Лемматизація, як нам здається вирішує цю проблему більш відповідним способом-приведенням слів до первісної форми (леми): називний відмінок в однині для іменників, інфінітив для дієслів, дієприслівників і т.д. хоча в будь-якому випадку, неможливо підібрати ідеальний спосіб подібного роду підготовки і тому змиритися з певними втратами все ж необхідно.

Наступним етапом перед підготовки тексту є виділення так званого «стоп-словника». Реалізація стоп-словників може відбуватися кількома шляхами:

– реалізація за допомогою видалення заданих слів, включених до словника (який можна розширити при бажанні термінами з окремих галузей);

– реалізація за допомогою виокремлення з результуючого списку-слів, які мають рівномірну зустрічальність в межах аналізованого масиву текстів і, отже, не становлять значущості в статистичному аналізі.

Як нам здається, обидва методи не є абсолютно придатними під завдання, що вирішуються розглянутим типом аналізу. В рамках першого типу губляться анафоричні посилання, виражені займенниками (що відсилають нас до вищезгаданих «суб'єктів» і «об'єктів» тексту). Другий спосіб в тій же мірі нівелює рівномірно вживані терміни, які, в свою чергу, можуть мати високий рівень значущості і включеності в різні пояснювальні структури. Втім, виходом з такої ситуації, як нам здається, може бути застосування стоп-словника в різних формах на різних етапах аналізу. Тобто. застосування класичної форми "стоп-словника «при підготовці корпусу текстів і подальше застосування стоп-словника при отриманні матриці термів-документів, для видалення слів, що вживаються раз, тільки в окремих документах корпусу, а також до семантично "малозначущих" одиниць.

Окрім того, при обробці корпусу текстів, нам вбачається важливим застосування TF-IDF показника. TF-IDF – є статистичною мірою, яка дозволяє визначити ваговий коефіцієнт для кожного терму в документі. Кожен терм підраховується в контексті його представленості (зустрічаємості) як в окремому документі, так і в корпусі аналізованих документів[3]. Дана статистична міра в більшості використовується з метою класифікації документів. У нашому дослідженні вона з одного боку дозволить доповнити «стоп-словник», визначивши терми, які в досить вираженому ступені зустрічаються у всіх документах корпусу, а з іншого може бути використана для підвищення ваги термінів, що використовуються в документах, але мають більш низький загальний рівень вживаності. Окрім того, TF-IDF не представляє великого інтересу на рівні аналізу корпусу з чотирьох звернень, які організовані за одним зразком і використовують ідентичну «жанрову» лексику. Але, при роботі

[1] Вакуленко, М. (2015). *Українська термінологія: комплексний лінгвістичний аналіз*. Івано-Франківськ: Фоліант.

[2] Штерн, І. (1998). *Вибрані топіки та лексикон сучасної лінгвістики: енциклопедія, словник*. Київ: Артек.

[3] Оськина, К. (2016). Оптимизация метода классификации текстов, основанного на TF-IDF, за счет введения дополнительных коэффициентов. *Вестник Московского государственного лингвистического университета*, 175-185.

з відкритим масивом документів, це дозволить виокремити дану лексику, виділивши її в якості жанрової особливості, без включення в кінцеву дискурсивну структуру.

На другому етапі аналізу тексту відбувається саме робота в напрямку виділення структуруючих знаків дискурсу. На різних етапах дослідження алгоритм виділення ключових знаків нам бачився по-різному, але його кінцева суть зводилася до того, що ми описували вище, а саме побудови мережі слів, здатних виражати дискурсивні підстави наявної системи. У процесі розробки і перевірки алгоритму автоматичного виділення дискурсивної структури з корпусу текстів (на прикладі звернень Президента України до ВРУ) нами були виявлені як методологічні, так і технічні недоліки, розроблюваного алгоритму, втім, вони в достатній мірі пов'язані, щоб описувати їх спільно.

Отже, хотілося б відзначити-це неможливість отримання термінологічної мережі з тексту безпосередньо. У будь-якому разі участь у визначенні того чи іншого знака у якості вузлової точки ставала дослідницьким рішенням, а це, з одного боку, підриває суть формального автоматичного аналізу, а з іншого дає нам результати з низьким рівнем верифікованості (детальніше про це можна подивитися в інших наших публікаціях..). Крім того, нам не вдалося отримати від тексту досить структуровану ієрархічну мережу, як ми припускали спочатку. Замість цього ми отримували ряд категорій, прямий зв'язок підпорядкування серед яких встановити досить складно (якщо не редукувати все до найпростішого рівня).

Окрім того, важливим є елемент контекстуального наповнення тих, або інших термінів. Тут ми маємо на увазі як усталені вирази, так і відносно «нейтральні» слова, що знаходять сенс в зв'язці з іншим термом. Даний рівень текстової (дискурсивної) організації можливо розглядати і аналізувати на рівні колокацій (словосполучень). В рамках проведеного дослідження звертати увагу варто було б на набір слів, які мають досить високий (вище певного порогу) рівень уживаності разом, що свідчить про деяку ступінь супідрядності і взаємопов'язаності (наприкл., Боротьба з корупцією, російська агресія тощо). Безумовно важливими є і анафоричні посилання на попередні слова тексту, реалізовані за допомогою займенників і займенникових слів[1].

Висновки: таким чином, формальний аналіз політичного дискурсу все ж вбачається можливим, проте потребує доробки першочергово на стадії епістемологічної відповідності між категоріями пошуку та можливостями технічної реалізації. Подальший розвиток алгоритму планується в сторону приведення категоріального апарату у відповідність з програмними можливостями автоматичного статистичного аналізу. А також продуктивним на даному етапі здається використання латентно-семантичного аналізу (lsa)[2] для побудови та уточнення прихованих внутрішньоструктурних відносин між термами текстового корпусу. Окрім того, розширення застосування того, що ми маємо на увазі під колокаціями (не варто плутати із сталими зверненнями (які в свою чергу все ж можуть представляти певний інтерес) та конвенційно і інституційно встановленими потребами в формулюваннях, зверненнях). Також, важливою вбачається ідея створення словника, що включає досить певну синонімію (сюди включені слова, які на досить очевидному рівні заміщають один одного і відносяться до одного і того ж внутрішньо текстового суб'єкту).

References:

1. Zhizhek, S. (2009). *Vozvyshennyj obekt ideologii* [The Sublime Object Ideology]. Moscow: Hudozhestvennyj zhurnal. [in Russian].
2. Howarth, D., Torfing, J. (editors) (2004). *Discourse theory in European politics: identity, policy, and governance*. Palgrave Macmillan: Basingstoke. [in English].
3. Bolshakova, E., Vorontcov, K., Efremova, N., Klyshinskii, E., Lukashevich, N., Sapin, A. (2017). *Avtomaticheskaia obrabotka tekstov na estestvennom iazyke i analiz dannykh* [Automatic Natural Language Processing and Data Analysis]. Moscow: NIU VShE. [in Russian].
4. Teodorescu, M. (2017). Machine Learning Methods for Strategy Research. *HBS Working Paper 18-011*. Harvard Business School. [in English].
5. Vakulenko, M. (2015). *Ukrainska terminolohiia: kompleksnyi linhvistychnyi analiz* [Ukrainian terminology: complex linguistic analysis]. Ivano-Frankivsk: Foliant. [in Ukrainian].
6. Shtern, I. (1998). *Vybrani topiky ta leksykon suchasnoi linhvistyky: entsyklopediia, slovnyk* [Selected topics and lexicon of modern linguistics: encyclopedia, dictionary]. Kyiv: Artek. [in Ukrainian].

[1] Большакова, Е., Воронцов, К., Ефремова, Н., Клышинский, Э., Лукашевич, Н., Сапин, А. (2017). *Автоматическая обработка текстов на естественном языке и анализ данных*. Москва: НИУ ВШЭ.

[2] Landauer, Th.W., Foltz, P., Darrell, L. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

7. Oskina, K. (2016). Optimizatciia metoda klassifikatcii tekstov, osnovannogo na TF-IDF, za schet vvedeniia dopolnitelnykh koeffitcientov [Optimisation of TF-IDF text classification method by introducing additional weighting coefficients]. *Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta* [Bulletien of Moscow State Linguistic University. Humanities], 175-185. [in Russian].
8. Landauer, Th. W., Foltz, P., Darrell, L. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *vol. 25*, 259–284. [in English].
9. Polevoi, N. (2010). *Prognozirovanie smeny paradigm istoricheskogo poznaniia: poisk metoda* [Forecasting the change of paradigms of historical knowledge: search for a method]. Odesa: Feniks. [in Russian].