

УДК 81'276.5

DOI: 10.31891/2415-7929-2021-21-22

ТАРАН О. С.

Національний університет «Львівська політехніка»

ЛЕБЕДЕНКО Ю. М.

Харківський національний педагогічний університет імені Г. С. Сковороди

СЕМАНТИКО-ДИСТРИБУТИВНИЙ АНАЛІЗ СЛЕНГІЗМІВ НА ОСНОВІ ДАНИХ КОРПУСУ ГРАК

У статті на прикладі сленгової лексики «дерибан» показано корпуснобазований підхід до семантико-дистрибутивного аналізу з метою виявлення латентних сем через оцінку регулярності колокацій. В основу дослідження покладено ідеї дистрибутивної семантики, що описує зміни в значеннях через частотні характеристики дистрибуцій у корпусі. Ключові слова: корпус, частотність, дистрибуція, сленгізм, MI, logDice.

TARAN O.

Lviv National Polytechnic University

LEBEDENKO YU.

Kharkiv National Pedagogic University named after H. Skovoroda

SLANG UNITS SEMANTIC AND DISTRIBUTIVE ANALYSIS WITH GRAK CORPUS DATA

The article describes a corpus-based approach to semantic and distributive analysis of Ukrainian slang unit "deriban". The study is based on the idea of distributive semantics which deals with semantic changes due to frequency of distributions in a corpus of texts. General Regionally Annotated Corpus of Ukrainian (GRAC) was taken as data base because it is the biggest Ukrainian corpus. It contains more than 728 million tokens (v. 11), also it has a linguistic annotation: morphological and semantic. The analysis of dictionary definitions revealed two homonymous words of different origins: the first one is from drug jargon, the second one is from criminal argot. But analysis of collocations with the word "deriban" in the corpus showed the prevalence of argot origin word in various functional styles. There are two spelling variants which are revealed in the sociolinguistic dictionaries but the corpus showed a high frequency of the one variant only. The corpus allows to identify the date of the first fixation of the word, as well as the frequency distribution by year. The slang unit semantic and distributive analysis with GRAK corpus data allows revealing both preserved archiseme of the argot unit and actualized semes and also latent semes. Discussion based on word occurrence frequencies, statistical parameters of slang unit collocation with MI and logDice scores. The result shows a stable collocation with the concepts of "land", "budget", which are obviously the most subject to division. Also it revealed some latent semes. T-score using shows the regularity of slang word spelling in quotation marks, which indicates the awareness of it as a substandard unit.

Key words: corpus, frequency, distribution, slang units, MI, logDice.

Постановка проблеми. Сленговий шар лексики є швидкозмінним: якісь одиниці виходять з активного ужитку, а якісь, отримуючи певні семантичні зміни, потрапляють в узус, зокрема розмовний шар літературної мови (так сталося, наприклад, з колишнім арготизмом «халтура»). Семантичні зміни відбуваються під впливом частотності входження в колокації, релевантне дослідження яких уможливило мовний корпус. Для цієї розвідки було обрано Генеральний регіонально анотований корпус (ГРАК) [7] як найбільший (понад 728 млн токенів) і з найповнішою метарозміткою, а також морфологічною й семантичною розміткою, що дозволяє робити різновекторні пошуки.

Аналіз останніх джерел. В основу цього дослідження покладено ідеї дистрибутивної семантики, що описує зміни в значеннях через частотні характеристики дистрибуцій у корпусі. Застосування класичних методів хвибує включенням лиш якоїсь вузької частини мовного матеріалу, суб'єктивізмом дослідника й відсутністю статистичної верифікації. Останнім часом спостерігаємо застосування методів комп'ютерної обробки даних, статистичного аналізу й моделювання семантичних змін. Переважно ці методики описані в зарубіжних наукових дослідженнях: дистрибутивні семантичні моделі для виявлення семантичних кластерів слів із семантичним зсувом у давньогрецькій мові [3], статистичні алгоритми для виявлення значущих семантичних зсувів в англійській мові на матеріалі трьох різних Інтернет-жанрів [2], метод латентного семантичного аналізу для виявлення суттєвих семантичних зсувів в історії англійської мови — базується на ідеї, що тенденція лексики з'являється в подібних контекстах є показником семантичної близькості [4]. У лінгвостатистичних дослідженнях зазвичай використовують такі показники колокацій: MI, T-score, logDice. Показник T-score включає до колокацій стоп-слова (займенники, прийменники, частки тощо, які є найчастотнішими), а також знаки пунктуації, лапки, а тому без стоп-списку не є показовим для з'ясування стійкості словосполучення [1, с. 348], однак він допоміжний у виявленні тенденції написання сленгізму в лапках. На показник MI сильно впливають низькочастотні слова, тому їх треба виключати. Перевага критерію logDice в тому, що він оцінює типовість колокацій, незалежно від розміру корпусу [5]. Корпусне програмне забезпечення NoSketchEngine, яким послуговується ГРАК, дозволяє обчислювати всі ці показники, з яких для

оцінки регулярності колокацій було обрано МІ з виключенням низькочастотних лем і logDice (формули розрахунку див: [5]).

Мета дослідження — на прикладі сленгової лексики «дерибан» показати корпуснобазований підхід до семантико-дистрибутивного аналізу з метою виявлення латентних сем через оцінку регулярності колокацій.

Виклад основного матеріалу.

1. Аналіз словникових дефініцій. «Словник української мови» в 20 томах фіксує лексему «дерибан» зі стилістичною позначкою «жаргонне» й таким тлумаченням: «Незаконний поділ, привласнення земель, майна, коштів і т. ін.» [11]. «Великий електронний словник української мови» фіксує лексему «дерибан» із семантичним тегом «сленгове» [6]. Тлумачні словники В. Бусела (2002, 2005) та 11-томний (1970–1980) цю лексему не фіксують, а отже, це непрямо вказує на її непоширеність у мовленні в ХХ ст. Цікаво, що в наївній картині світу, яку відбиває онлайн-словник МуСлово, лексема «дерибан» пов'язується з українським політикумом: «новий тренд української політики, що прийшов на зміну прихвизації. Полягає у розтягуванні майна держави чи громади або розподілі посад, які сидять на грошових потоках» (фіксація 2012 р.) [8].

Сучасні соціолектні словники української мови фіксують лексему «дербан» / «дерибан» як одиницю наркожargonу в таких значеннях: «збір урожаю маку» [13, с. 120; 9, с. 185], «пора цвітіння маку» [10, с. 108; 9, с. 185], «сезон збирання маку», «процес збирання маку» [10, 108]. «Словник луцьких жаргонізмів і сленгізмів» Г. Аркушина (2011) цю лексему не засвідчує. Натомість у цих же працях знаходимо дієслово «дерибанити» / «дербанити», дотичне як до жаргону наркоманів, так і кримінального аргю в значенні «поділ награваного», що свідчить про змішування омонімів. Поширення «дерибан» як одиниці з наркожargonу чи кримінального аргю дозволяє з'ясувати аналіз корпусних даних.

Українські соціолектні словники кінця ХІХ – початку ХХ ст. (насамперед це праці О. Горбача) аналізовану лексему не фіксують. Натомість її виявлено в лексиконах російської мови, складених працівниками карного розшуку для вузькопрофесійних потреб, у зв'язку з чим вони тривалий час мали обмежений доступ. На початку ХХ ст. трапляються фіксації переважно лексем рос. *дербанка* — «частка» (Г. Досталь, «Воровской словарь», Слобода Покровская, 1904), «поділ» (В. Лебедев, «Словарь воровского языка», 1909; «Арестантский словарь», 1913; Н. Хандзинский, «Словарь блатного жаргона», 1926), «поділ, хабар» («Блатная музыка. Словарь жаргона преступников», Москва, 1923), «поділ; хабар; подаяння в тюрмі» (С. М. Потапов, «Словарь жаргона преступников (Блатная музыка)», Москва, 1927), «поділ награваного» (В. Каверин, «Слова воровского языка, встречающиеся в повести «Конец хазы», Ленинград, 1930); рос. *дербанщик* — «крадій, який бере дербанку» (В. Лебедев, «Словарь воровского языка», 1909), «учасник зі сторонніх» (В. М. Попов, «Словарь воровского и арестантского языка», Киев, 1912; «Блатная музыка. Словарь жаргона преступников», Москва 1923; С. М. Потапов, «Словарь жаргона преступников (Блатная музыка)», Москва, 1927) [12]. Лексема рос. *дербанить* регулярно фіксується словниками, починаючи з середини ХХ ст.

2. Лексема «дерибан» крізь призму статистики текстів корпусу.

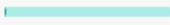
Lemma	↓	Frequency	Frequency per million
1 дерибан		723	0.99  ...
2 дербан		2	< 0.01  ...

Рис. 1. Частотність варіантів лексики

Як бачимо з рис. 1, первинний варіант *дербан* не набув поширення в українській мові. Оскільки його частота не є статистично значущою, надалі досліджуватимемо дистрибуції лексики «дерибан».

Загальна кількість виявлених словоформ (токенів) лексики «дерибан» — 726. Сортювання отриманого конкордансу за метаданими дає змогу побачити, зокрема, частотність входжень у тексти відповідних стилів. Найбільшою мірою ця лексема трапляється в публіцистичних текстах (611), суттєво рідше в усно-розмовному стилі (82), далі за спадною: художніх текстах (17), его-текстах (щоденники, листи, блоги, спогади, автобіографії, 13) і 3 фіксації в науковому стилі як ілюстративний матеріал — у подальших запитах цей ресурс було виключено. Ці результати залежні від ступеня наповнення корпусу, тобто від кількості токенів у кожному стилі: художній і публіцистичний мають зіставлювані величини, а розмовний — у 7 разів менший, як і его-тексти.

Сортювання за джерелами дозволяє виявити, що найактивніше вживають лексему «дерибан» Інтернет-газети «Україна молода», «Високий замок», «День», «Фіртка». Електронні ЗМІ сьогодні посідають перше місце за популярністю серед споживачів інформації, тож вони є найпотужнішими інструментами впливу на ціннісну картину світу й мовні смаки споживачів інформації.

CQL-графік (рис. 2) показує частотний розподіл лексеми за роками (на 1 млн), а також першу фіксацію в корпусі — 1993 рік.

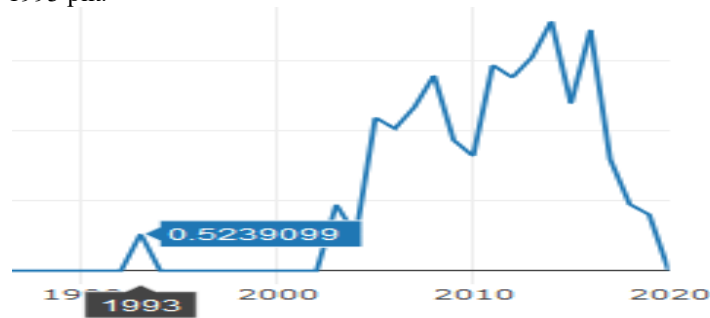


Рис. 2. Частотність лексеми «дерибан» за роками

3. Об'єктні дистрибуції лексеми «дерибан» описує модель «дерибан+іменник у Р.в.», що зреалізована в ширшому CQL¹-запиті [lemma="дерибан"][tag="noun.*"]. Після ручного виключення біграм, що не відповідають заданій моделі (як-от: *земельний дерибан очільника обласного комітету земельних ресурсів; віддають на дерибан можновладцям*) було отримано конкорданс зі 159 контекстів (частотність 0,22 на 1 млн), який посортовано насамперед за частотністю правих дистрибуцій першого рівня (R1). У підрахунку також було враховано дистрибуції R3,4,5, якщо вони відповідали заданій моделі (як-от: *дерибан військового майна; дерибан системи охорони здоров'я; дерибан сотень гектарів дорогої землі*). Семантично близькі поняття погруповано за ЛСГ (одиничні поняття, не співвідносні з жодною ЛСГ, до уваги не брали), частотний їх розподіл відносно загальної кількості отриманих контекстів (159) такий:

ЛСГ «Земля» (42%): земля: активний дерибан земель, для демонстрації зухвалою і безкарного дерибану земель.

ЛСГ «Кошти» (19%): бюджет, держбюджет, кошти, гроші, мільярди, активи, субвенції, бабки, обшак: солодкий дерибан бюджету; дерибан мільярдів «Укроборонпрому»; дерибан коштів під час будівництва об'єктів інфраструктури; можливість дерибану всіх державних активів; історія з дерибаном субвенцій; черговий дерибан грошей; дерибан бабок; дерибан обшака; винних у багатомільйонному дерибані держбюджету.

ЛСГ «Країни, макрорегіони» (9%): країна, Україна, територія України, Крим, півострів, Польща, Чехословаччина: провести дерибан Криму; взяти участь у дерибані Польщі; ганебну участь у дерибані Чехословаччини; прикриває дерибан країни; відпочивати від трудів праведних по дерибану України; щодо дерибану території України; дерибану всього півострова.

ЛСГ «Посади» (9%): посади, портфелі: розпочинається дерибан посад; дерибан портфелів.

ЛСГ «Нерухоме майно» (8%): майно, нерухомість, приміщення, власність, шахти: дерибан чужого майна; дерибан нерухомості; дерибан приміщень; технології дерибану власності; дерибан росіянами шахти.

ЛСГ «Міста, урбаністичні об'єкти» (6%): Київ, Коломия, парки, сквери: нова влада розпочала дерибан Коломиї; плакати із закличками «Не допустимо дерибану Києва»; голосування щодо дерибану парків і скверів.

ЛСГ «Природні ресурси» (3%): ліс, ресурс, зелені інвестиції: дерибан лісу; дерибану останнього рекреаційного ресурсу громади; дерибан зелених інвестицій.

ЛСГ «Влада» (3%): влада, сфера впливу: дерибан влади; дерибану сфер впливу.

ЛСГ «Електорат» (1%): електорат, округ: дерибан електорату; нову схему дерибану округів на виборах 2012.

Як бачимо, поділу підлягають великі і/або дороговартісні об'єкти: лексема «дерибан» не сполучається з лексемами на позначення чогось дрібного чи дешевого. Крім того, серед 159 контекстів не було виявлено жодного, який би експлікував значення, детерміноване наркожаргоном, що засвідчене соціолектними словниками.

В оцінці регулярності сполучуваності релевантними вважають показники MI, logDice >1. Для MI обрано мінімальну абсолютну частоту лема 5. Серед дистрибутивів R1 найвищі показники за обома критеріями мають лема *нерухомість* (MI=8.16, logDice=3.13), *посада* (MI=7.70, logDice=2.74), *земля* (MI=7.48, logDice=2.53), *бюджет* (MI=6.63, logDice=1.67), *кошт* (MI=6.34, logDice=1.39).

4. Атрибутивні дистрибуції лексеми «дерибан» описує модель «прикметник +дерибан», яка зреалізована в CQL-запиті [tag="adj.*"&tag!="*pron.*"][lemma="дерибан"]. У результаті отримано 144 контексти з 82 атрибутивними лемами, що вказують на: об'єкт дерибану (*земельний, бюджетний, трубний, фінансовий, тендерний, рибний, приватизаційний, пляжний, парламентський, партійний, кадровий, паркувальний, мілітарний, лісовий, земельно-лісовий, екологічний, грошовий, господарчий, владний,*

¹ CQL — це мова запитів у корпусі (від Corpus Query Language).

бюджетно-податковий), масштаб (великий, грандіозний, найбільший, широкомасштабний, масовий, загальний, космічний, вічний, всезагальний, багатомільярдний, багатомільйонний, безпрецедентний, величезний), спосіб здійснення (підкилимний, підпільний, інтенсивний, тихий, спільний, попередній, неприхований, неприкритий, запланований, договірний, безперешкодний, безсистемний), регулярність (черговий, остаточний, новий, наступний), загальну та емоційну оцінку (цинічний, солодкий, фесричний, справжній, банальний, нечуваний, переможний, нищівний, нахабний, зухвалий, жорстокий, ефективний, безсовісний, ганебний, відвертий), відношення до закону (безкарний, злочинний).

Оцінка сполучуваності лексеми «дерибан» на першому рівні (L1) виявила такі прикметники: підкилимний (MI=12.94, logDice=6.47), безкарний (MI=11.91, logDice=6.03), земельний (MI=9.37, logDice=4.40), банальний (MI=9.33, logDice=4.23), черговий (MI=7.18, logDice=2.23), бюджетний (MI=6.78, logDice=1.80), а також високу сполучуваність з дієсловами зупинити (MI=6.85, logDice=1.88), припинити (MI=7.42, logDice=2.44) та іменниками схема (MI=7.92, logDice=2.95), масштаб (MI=7.03, logDice=2.04). Як бачимо, поняття «земля», «бюджет» як об'єкти розподілу, пограбунку сполучаються з лексемою «дерибан» і як атрибутивні (частіше), і як об'єктні.

Критерій T-score засвідчує найвищу з-поміж усіх колокацій частотність поєднання ключової лексеми з лапками, а це вказує на усвідомлення авторами текстів ненормативності лексеми.

Висновки. Корпуснобазований підхід до аналізу сленгізму «дерибан» виявив: 1) поширеність запозиченої з російського кримінального аргю лексеми в різних функційних стилях, тоді як її омонім з наркожаргону, фіксований українськими соціолектними словниками, так і не вийшов за межі соціолекту; 2) поширеність лексеми «дерибан» попри первинність варіанта «дербан» у мові-джерелі; 3) запозичення, очевидно, відбулося на початку 90-х років, що збігається з часом чергової розгерметизації аргю; 4) найпотужнішим джерелом функціонування лексеми «дерибан» є електронні ЗМІ; 5) аналіз правих і лівих дистрибуцій виявив збережену архісему «розподіл», актуалізовані семи «незаконний», «безкарний», деактуалізацію через функційне розширення семи, що вказувала на соціальне обмеження функціонування арготизму, актуалізацію об'єктної семи «дороговартісний» та латентних іманентних сем «можновладець як суб'єкт дії», а також негативну конотацію; 6) оцінка колокацій лексеми «дерибан» виявила стійку сполучуваність з поняттями «земля», «бюджет», які, очевидно, найбільшою мірою підлягають поділу; 7) високі показники MI, logDice для дистрибутивів L1 дозволяють вказати ще на такі актуалізовані латентні семи: «масштабність незаконних розподілів», «регулярність», «схематичність», а також на семи, що описують суспільну відповідь «необхідність припинення»; 8) оцінка за T-score виявила регулярність написання слова в лапках, що свідчить про усвідомлення його як субстандартної одиниці.

Отже, усе це свідчить про входження лексеми «дерибан» до загального сленгу, одиниці якого добре відомі всім членам мовного колективу, незалежно від соціальної його стратифікації. Крім того, такі сленгізми проникають і в ономастичний простір (пор.: *новий пам'ятник «Дерибану Києва»*).

Література

1. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: корпусные подходы / под ред. А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой. Хельсинки, 2008. С. 343–357.
2. Kulkarni V., Perozzi B., Al-Rfou R., Skiena S. Statistically Significant Detection of Linguistic Change. 2014. URL: <https://arxiv.org/abs/1411.3315>.
3. Rodda M. A., Senaldi M. S. G., Lenci A. Tracking Semantic Change with Distributional Semantics in Ancient Greek // Italian Journal of Computational Linguistics (IJCoL). 2017. № 3–1. P. 11–24. DOI: 10.4000/ijcol.421.
4. Sagi E., Kaufmann S., Clark B. Tracing semantic change with Latent Semantic Analysis // Current Methods in Historical Semantics / edited by Kathryn Allan and Justyna A. Robinson. Berlin, Boston: De Gruyter Mouton, 2011. P. 161–183. DOI: 10.1515/9783110252903.161
5. Statistics used in the Sketch Engine. Lexical Computing Ltd, 2015. URL: <https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine>

Джерела

1. Великий електронний словник української мови (версія 5.3.1, 2021) / А. Рисін, В. Старко, команда БрУК. URL: <https://r2u.org.ua/vesum>
2. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко, М. Возняк, М. Крук та ін. Київ, Львів, Єна, 2017–2021. URL: uacorporus.org.
3. MyСлово: словник, який пишеш ти. URL: <http://myslovo.com/>
4. Пиркало С. Перший словник українського молодіжного сленгу. Київ, 1997.
5. Словник сучасного українського сленгу / упор. Т.М. Кондратюк. Харків: Фоліо, 2006.
6. Словник української мови: у 20 т. URL: <http://lcorp.ulif.org.ua/ExplS/>
7. Собрание толковых словарей тюремного и блатного жаргона: 26 словарей, изданных в России с

1859 до 2005 гг. / под ред. Е. Ю. Ачеркан. CD-диск (программная оболочка ETS), 2005.
8. Ставицька Л. Український жаргон. Київ: Критика, 2005.

References

1. Hohlova M.V. Eksperimental'naya proverka metodov vydeleniya kollokacij // Slavica Helsingiensia 34. Instrumentarij rusistiki: korpusnye podhody / pod red. A. Mustajoki, M.V. Kopoteva, L.A. Biryulina, E.Yu. Protasovoj. Hel'sinki, 2008. S. 343–357.
2. Kulkarni V., Perozzi B., Al-Rfou R., Skiena S. Statistically Significant Detection of Linguistic Change. 2014. URL: <https://arxiv.org/abs/1411.3315>
3. Rodda M. A., Senaldi M. S. G., Lenci A. Tracking Semantic Change with Distributional Semantics in Ancient Greek // Italian Journal of Computational Linguistics (IJCoL). 2017. № 3–1. P. 11–24. DOI: 10.4000/ijcol.421.
4. Sagi E., Kaufmann S., Clark B. Tracing semantic change with Latent Semantic Analysis // Current Methods in Historical Semantics / edited by Kathryn Allan and Justyna A. Robinson. Berlin, Boston: De Gruyter Mouton, 2011. P. 161–183. DOI: 10.1515/9783110252903.161.
5. Statistics used in the Sketch Engine. Lexical Computing Ltd, 2015. URL: <https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine>

Sources

1. Velykyi elektronnyi slovnyk ukrainskoi movy (versia 5.3.1, 2021) / A. Rysin, V. Starko, komanda BrUK. URL: <https://r2u.org.ua/vesum>
2. Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK) / M. Shvedova, R. Fon Valdenfels, S. Yaryhin, A. Rysin, V. Starko, M. Vozniak, M. Kruk ta in. Kyiv, Lviv, Yena, 2017–2021. URL: uacorp.us.org.
3. MySlovo: slovnyk, yakyy pyshesh ty. URL: <http://myslovo.com/>
4. Pyrkalo S. Pershyi slovnyk ukrainskoho molodizhnogo slenhu. Kyiv, 1997.
5. Slovnyk suchasnoho ukrainskoho slenhu / upor. T.M. Kondratiuk. Kharkiv: Folio, 2006.
6. Slovnyk ukrainskoi movy: u 20 t. URL: <http://lcorp.ulif.org.ua/ExplS/>
7. Sbranie tolkovykh slovarej tyuremnogo i blatnogo zhargona: 26 slovarej, izdannyh v Rossii s 1859 do 2005 gg. / pod red. E.Yu. Acherkan. CD-disk (programmaya obolochka ETS), 2005.
8. Stavytska L. Ukrainskyi zhargon. Kyiv: Krytyka, 2005.