81’25’42=81'322

***С.Б.Фокін, доцент, канд.філол.н.,***
***КНУ імені Тараса Шевченка***

# ПЕРЕВАГИ КОРПУСІВ МАЛОГО ОБСЯГУ ДЛЯ ДОСЛІДЖЕННЯ МАКРОЯВИЩ У ПЕРЕКЛАДІ

*Достатній обсяг корпусу може варіюватися залежно від потреб дослідження. Під час дослідження макроявищ в оригінальних та перекладних текстах принципово треба розмежовувати такі що 1) не піддаються формалізації; 2) що формалізуються частково; 3) що не формалізуються. Для перших і других доречно використовувати корпуси малого обсягу, які в цьому випадку мають низку переваг.*

***Ключові слова:*** *переклад, макростилістика, корпусна лінгвістика, лінгвостатистика.*

У перекладознавчих дослідженнях дедалі частіше використовується кількісний аналіз, який, у поєднанні з досягненнями корпусної лінгвістики, здатний давати часом несподівано цікаві результати. Втім, на терені українського перекладознавства необхідно враховувати певні поправки. По-перше, корпусною лінгвістикою, надбанням, що прийшло з Заходу, володіють радше вибрані вчені. По-друге, що є наслідком першого, чимало комп’ютерних знарядь іще не розроблені для української мови й, тим паче, для пар іноземна мова/українська. І хоча для української мови наявні корпуси з потужним знаряддям пошуку [Корпус 2014], [Український 2014], усе ще відсутній доступний лематизатор, комп’ютерний тезаурус та інші корисні знаряддя. Тому в рамках часткових теорій перекладу дослідники, що прагнуть показати кількісний або навіть статистичний аспект досліджуваного об’єкта, використовують власноруч створені корпуси. На перший погляд це незручно й для дослідника, якому самотужки доводиться опрацьовувати великі масиви текстів. До того ж, зібраний *ad hoc* дослідницький матеріал поступається своїм обсягом. Втім, менша кількість не обов’язково означає гіршу якість: у малих дослідницьких корпусів є низка переваг, розгляд яких є метою статті.

Безперечно, будь-які зусилля дослідника створити власну емпіричну базу затьмарюється кількісною перевагою потужних корпусів, розроблених на Заході. І все ж не можна заперечувати, що часом завеликий обсяг може тільки завадити дослідженню й його ускладнити. Крім того, на відміну від лінгвіста, перекладознавці працюють з кількома мовами паралельно. Для пари іспанська та українська мови, окрім зазначених українських корпусів, актуальним є використання Дослідницького корпусу української мови [Корпус 2014] та корпусу *CREA (Corpus de Referencia del Español Actual*

[Corpus 2014]) Іспанської королівської академії. Втім, згідно з характеристиками корпусів, вказаними на сайтах, Дослідницький корпус української мови нараховує 13 мільйонів слововживань, а *CREA* – 160 мільйонів. Очевидно, що для дослідника, який оцінює певне явище в обох мовах, у контрастивному плані така диспропорція виявиться незручною. Власноруч створити мільйонний корпус видається справою невдячною, якщо йдеться про уточнення якоїсь дрібної деталі або ж для реалізації первинного пошуку й виявлення загальних тенденцій. Л. Боукер слушно зауважує, що часто корпус, який укладається хіба що на один-два використання, не вартий значних часових затрат [Bowker 2001: 351-352]. А втім, спочатку пропонуємо дати відповідь, коли можна й треба обходитися без мільйонних корпусів, коли цілком достатньо обійтися скромнішим обсягом.

Для початку проаналізуймо, якого порядку обсяг матеріалу пропонують використовувати різні дослідники у філологічних науках взагалі та в перекладознавстві зокрема. Скажімо, укладачі “Частотного словника української мови” і “Оберненого частотного словника української мови” – Т.О. Грязнухіна, Н.П. Дарчук, Є.А. Карпіловська, Н.Ф. Клименко, Л.І. Комарова, В.І. Критська, Л.В. Орлова [Частотний 1981]. У зв’язку з цим Є.А. Карпіловська відзначає: “Чому п’ятсот тисяч, – спитають мене ті, хто знає, що нині корпуси оперують уже обсягами у сотні мільйонів слововживань. А тому що лінгвостатистика уже давно довела, що не треба займатися для виконання багатьох завдань гігантоманією, а цілком досить для того, щоб одержати вірогідні характеристики частоти вживання слова в тексті – достатньо вибірки не меншої в 300 тис. слововживань. І будуть надійними ті статистичні характеристики, які ми одержали. Отже, текстового обсягу у 500 тис. слововживань було цілком вдосталь, для того щоб визначити частоту вживань того чи іншого слова в нашому реєстрі” [Карпіловська 2011]. Додамо, що це не суперечить доцільності укладання великих мільйонних корпусів. Зокрема, в них “користувач може відібрати об’єкти збереження, які його цікавлять, з різних пошукових запитів (…) користуючись таким інструментом, дослідник може відібрати джерела певного автора, стилю чи жанру і працювати лише з цією частиною корпусу, фактично ця процедура дає можливість виділити із загального корпусу свій власний підкорпус, орієнтований на розв’язання особистих завдань” [Широков 2011: 158]. Наприклад, можна проводити пошук не в усьому корпусі, а виключно в газетно-журнальних статтях. Однак навіть у таких корпусах зі змінними параметрами є маса обмежень. Якщо, скажімо, дослідника цікавлять наукові

статті з психології, а в діалоговому вікні можна вибрати лише “періодику” і “природничі науки”, недостатня гнучкість параметрів корпусу теж не вирішить проблему збору матеріалу. Тому самостійний підбір матеріалу часто лишається єдиним можливим засобом забезпечити емпіричну базу дослідження. При сучасних можливостях комп’ютерів цілком можливо довести цю базу й до обсягу у 500 тисяч слововживань. Але чи завжди потрібний саме такий обсяг?

Коли ідеться про достатність півмільйонного корпусу, мається на увазі створений незмінюваний корпус для певного дослідження, як і було зроблено для укладання указаних частотних словників. Вони репрезентують українську художню прозу взагалі. Натомість, скажімо, досліднику народних казок буде потрібно складати окремий корпус текстів досліджуваного жанру, а досліднику чарівних казок – інший. Питання полягає в тому, чи потрібно досліднику прагнути обсягу 300.000-500.000, коли корпус текстів укладається не для зрізу мови взагалі, а для певної вузької мети. Виявляється, є схвальні відгуки і про корпуси ще меншого обсягу, від 20.000 до 200.000 слововживань. Зокрема, Г.Інгвел у статті “Не випадковість, а свідомий вибір: критерії укладання корпусу” (“*Not chance but choice: criteria in corpus creation”)*, окрім загальних корпусів, розглядає корпуси спеціальні, в тому числі так звані “якісні корпуси”. “Якісний корпус (…) розробляється з метою ширшої і більш репрезентативної вибірки досліджуваної спеціалізованої мови (…). Дійсно, в корпусній лінгвістиці те, що більше, не обо’язково краще. Дрібний, однак добре організований корпус може виявитись кориснішим, ніж більший за обсягом, але менш обміркований за змістом (...)- За свої досвідом я переконався, що корпуси від 20 000 до 200 000 слововживань виявились цілком придатними для здійснення дослідження. Більші корпуси надто довго укладаються, а менші містять недостатньо даних для належної інтерпретації” [цит. за Bowker 2001, 352]. Розвиваючи думку вченого, зауважимо, що терміни як формально виокремлювані лексичні одиниці відносно легко піддаються формалізації навіть у флективних мовах. Тому їх можна легко й безперешкодно досліджувати і в гігантських національних корпусах, а підрахунки машина здійснює за лічені секунди. Перевага малого корпусу буде в тому, що той чи інший термін дослідник зможе переглянути на власні очі й відбракувати можливі омонімічні вживання. Натомість формальний автоматизований пошук у мільйонних корпусах не дає такої змоги.

Однак значно цікавіше поглянути на проблему вибірки фактичного матеріалу очима дослідників, які досліджують явища, що лише частково піддаються формалізації або не піддаються їй зовсім. Скажімо, майже не формалізується пошук метафор, хоча є спроба оцінки метафоричності тексту. Зокрема, хоч і опосередковано, кількісний підрахунок типових та нетипових колокацій у перекладному тексті може слугувати певною кількісною оцінкою його метафоричності. Дослідниця Б. Левандовська-Томашчик зосереджує увагу на лексико-семантичній сполучуваності іменників оригіналу на позначення емоцій [Lewandowska 2012: 20-21]. Зрозуміло, що, чим менш типовою є сполучуваність, тим більшою є метафоричність – приклад полською мовою – “кипіти від сорому” – “kipiec ze wstydu”. Однак пошук колокацій доволі легко формалізується, тому цілком виправдано, що дослідниця використала у функції конкордансу оригінальні й паралельні тексти з колосальним обсягом – від 10 до 65 мільйонів слововживань. Цікаві висновки й методологічні знаряддя дослідниці корисно поширити і для співставлення оригіналу й перекладу: чи перекладається нетипова сполучуваність аналогічно екстравагантною, свіжою метафорою; варто перевірити, чи кількість нетипових колокацій у першотворі й цільовому тексті приблизно однакова, чи разюче відрізняється. Безумовно, для такого дослідження бажано укладати репрезентативний, але водночас легкоосяжний для дослідника корпус.

Розглянемо тепер, коли малі корпуси можуть виявитися корисними для дослідження таких груп явищ: 1) що не піддаються формалізації; 2) що піддаються частково формалізації; 3) що формалізації піддаються повністю.

Прикладом **явищ, що майже не піддаються формалізації**, можуть бути дисфемізми. Сукупність вживання дисфемізмів у художньому творі задає ті чи інші конотації його складовим частинам, позаяк дисфемізми виражають негативну оцінку денотата мовцем. Тому принципово важливо відстежити, чи зберігається в перекладі сукупна конотація. Завдяки широкому використанню механізмів втрат і компенсацій, а в цьому випадку йдеться про стилістичні втрати і компенсації, важливо передати не стільки кожен дисфемізм дисфемізмом, скільки “конотативний тон” узагалі. Звичайно ж, успішність розв’язання цього завдання можна оцінити тільки з залученням кількісного аналізу в макроконтексті. В окремому дослідженні ми здійснили попередню первинну розвідку щодо використання дисфемізмів у двох розділах роману А.Переса Реверте “Шкіра для барабана” у перекладі О.Леська. В аналізованому фрагменті оригіналу виявилось близько 100

дисфемізмів, щоправда, лише половина з них була відтворена стилістично зниженими одиницями. Втім, у фрагменті перекладу теж ми нарахували близько 100 дисфемізмів, тобто, на 50 втрачених стилістично знижених одиниць мовлення з’явилося приблизно 50 компенсацій у контексті. Таким чином, попри відсутність часом конотативної еквівалентності на мікрорівні, на макрорівні ця еквівалентність витримується [Фокін 2014: 464]. Цікавий цей висновок насамперед тим, що, безумовно, хоча перекладачі не проводять кількісні підрахунки дисфемізмів у першотворі і своєму перекладеному тексті, а при відтворенні конотації радше покладаються на інтуїцію, конотативна макроеквівалентність витримується. Звісно, ця попередня розвідка потребує глибшого аналізу на ширшому фактичному матеріалі, адже для глобальних висновків двох розділів роману замало. Однак оскільки дисфемізми практично не формалізуються (у них є певні не облігаторні маркери), вибірку може здійснювати виключно людина. Формальними підказками можуть бути маркери дисфемізмів, напр., вказівні займенники, означений артикль, відхилення від норм орфографії, суфікси зневажливої семантики – [Фокін 2014: 469], а також властивість дисфемізмів розташовуватися у тексті скупчено [Фокін 2014: 470]. Більше того, оскільки аналіз конотативного рівня безумовно пов'язаний з індивідуальним сприйняттям оцінки, вибірка може бути не позбавлена певної суб’єктивності. Тому тут навіть важливо витримувати не лише єдині критерії вибірки, а й щоб її з початку до кінця здійснювала одна й та сама людина. Подібні вимоги, до речі, висуваються до деяких медичних досліджень, коли результат дослідження залежить від інтерпретації лаборанта. Такі вимоги допускаються тоді, коли має значення не стільки абсолютна кількість, скільки динаміка розвитку цієї кількості (збільшення або зменшення). Так само і в випадку роботи з дисфемізмами значущим є не стільки бездоганна арифметична точність підрахунку дисфемізмів, скільки приблизна кількість і динаміка її збільшення або зменшення в перекладі.

З огляду на це, звичайно, покладатися на формальний пошук тут неможливо й немає сенсу. Оскільки вибірка цілком лягає на плечі людини, то можна рекомендувати використовувати для початку мінімально рекомендовані корпуси у 20 000 слововживань. Для порівняння, у найдовшій п’єсі В.Шекспіра “Гамлет” нараховує приблизно 30 000 слів. Іншими словами, переклад твору, співвідносного за обсягом з “Гамлетом”, можна вважати вже в певному сенсі мірилом приблизно мінімального стартового корпусу, якщо не досліджується явище, що потребує сукупності текстів, які б

репрезентували мову взагалі. Для початку це могли б бути навмання відібрані фрагменти у 5 000 слововживань з чотирьох різних художніх творів. За умов досягнення у першому наближенні ствердних результатів емпіричну базу доцільно розширювати. Під час строгих математичних розрахунків для перевірки достовірності корпусу слід використовувати вже наявні лінгвостатистичні знаряддя [Носенко 1981: 69], [Перебийніс 1985: 57]. Практика свідчить, що матеріал порядку сотень тисяч слововживань показує надійні критерії достовірності [Фокін 2011: 207].

**Явища, що частково формалізуються.** В іншому окремому дослідженні ми проводили пошук герундіальних перифраз в іспаномовному перекладному тексті, які слід розглядати формальними мовними одиницями. Тим не менші, окрім власне того, що наявні знаряддя пошуку не дозволяли вести пошук двослівних комбінацій зі змінюваним першим словом, скажімо “fue creciendo” – програма з лематизатором мала б знайти усі форми дієслова “ir”. І навіть якби й була така можливість, в окремих випадках тільки в контексті можна з’ясувати, чи йдеться справді про перифразу, чи дієслово перед герундієм вжито у своєму прямому значенні (“va llorando”). Тож, корисно й доречно виявилося переглянути кожен приклад на власні очі.

Таким чином, новітні й несподівані результати можна отримувати за допомогою відносно невеликих корпусів текстів оригіналу й відповідних перекладів. Зокрема, тільки невеликий обсяг може опрацювати дослідник, коли явище не піддається формалізації, якщо вчений має самостійно прочитати й проаналізувати тексти. Однак трапляються випадки, коли на малому корпусі приголомшливо несподівані висновки отримають навіть для **явищ, що цілком формалізуються**. Зокрема, згадаймо дослідження А.Б.Кутузова (викладача Тюменського університету, РФ). Розраховуючи індекс TTR (частку від кількості словоформ на кількість слововживань, що опосередковано дозволяє виміряти ступінь лексичного багатства тексту), дослідник виявив, що індекс розбіжний в цільовому тексті й першотворі. Однак динаміка його змін від розділу до розділу становить напрочуд схожу картину і в оригіналі, і в перекладі [Kutuzov 2010]. Підкреслимо, що надзвичайно наочні й перспективні результати були отримані на матеріалі дослідження двох романів К. Воннеґута та їхніх перекладів.

Таким чином, за певних умов масштабний загальний дослідницький корпус не відповідає потребам дослідження й виявляється доцільним укладати корпус *ad hoc*. Зокрема, такі кроки слід вважати виправданими, якщо досліджуваний об’єкт не підлягає формалізації або підлягає їй лише

частково. У такому випадку вибірку й кількісні підрахунки може здійснювати виключно сам дослідник. Відтак, зримим завданням може бути опрацювання корпусу малого обсягу. У перекладознавчих дослідженнях порядок малих корпусів подекуди починається з 20 000 слововживань. На наш погляд, такий обсяг доцільний під час первинної розвідки або при з'ясуванні другорядних характеристик об'єкта.

При строгому розрахунку достатності корпусу за статистичними критеріям порядок кількох сотень тисяч слововживань виявляється достатнім. Водночас це такий обсяг, аналіз якого цілком здатний подужати дослідник у процесі написання дисертаційного дослідження. Оскільки обсяги сучасних корпусів рахуються мільйонами слововживань, вважаємо справедливим називати "малими" корпуси порядку сотень тисяч слововживань.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:


*1. Карпіловська Є.А.* Українська комп'ютерна лінгвістика сьогодні: суспільні замовлення – здобутки – проблеми [лекція] / Є.А. Карпіловська. – [Електронний документ]. Режим доступу: http://polit.ua/lectures/2011/01/19/karpilovskaya.html

*2. Корпус* текстів української мови. – [Електронний документ]. Режим доступу: http://www. mova.info/corpus.aspx?l1=209

*3. Носенко И.А.* Начала статистики для лингвистов / И.А. Носенко. – М.: Высш. школа, 1981. – 156 с.

*4. Перебийніс В.С.* Частотні словники та їх використання / В.С. Перебийніс, М.П. Муравицька, Н.П. Дарчук. – К.: Наукова думка, 1985. – 204 с.

*5. Український* національний корпус. – [Електронний документ]. Режим доступу: http://lcorp.ulif.org.ua/virt_unlc/

*6. Фокін С.Б.* Дисфемізми в художньому перекладі: еквівалентність у мікро- та макроконтексті / С.Б. Фокін. // Мовні і концептуальні картини світу. – К.: ВПЦ "Київський університет", – Вип. 47. – Ч. 2. – С. 473-484.

*7. Фокін С.Б.* Маркери дисфемізмів у художньому тексті (на матеріалі іспанської та української мов) / С.Б. Фокін. // Проблеми семантики, прагматики та когнітивної лінгвістики. – К.: Логос, 2014. – Вип. 25. – С. 462-472.

*8. Фокін С.Б.* Критерій достатності корпусу при розрахуванні фактору частотності в перекладі / С.Б. Фокін. // Вісник Харківського університету: Романо-германська філологія. Методика викладання іноземних мов. – Харків: Харківський національний університет імені В.Н. Каразіна, 2011. – Вип. 68. – С. 200-207.

*9. Частотний* словник сучасної української художньої прози / за ред. Перебийніс В.С., редкол.: Н.П. Дарчук, Н.Ф. Клименко, В.І. Крітська, М.П. Муравицька, Л.В. Орлова. – К.: Наук. думка, 1981. – Т. 1 . – 863 с.

*10. Широков В.А.* Комп'ютрна лексикографія / В.А. Широков. – К.: Наукова думка, 2011. – 351 с.

*11. Bowker L.* Towards a Methodology for a Corpus Based Approach to Translation Evaluation / L. Bowker // Meta, 2001, 46, #2. – [Електронний документ]. Режим доступу: http://www.erudit.org/revue/meta/2001/v46/n2/002135ar.pdf

*12. Corpus* de Referencia del Español Actual. – [Електронний документ]. Режим доступу: http://www.rae.es/recursos/banco-de-datos/crea

*13. Kutuzov A.* Change of word types to word tokens ratio in the course of translation (based on Russian translations of K. Vonnegut's novels) / A. Kutuzov // International Computational Linguistic Conference "Dialog-21" – [Електронний документ]. Режим доступу: http://arxiv.org/ftp/arxiv/papers/1003/1003.0337.pdf

*14. Lewandowska-Tomaszczyk B.* Explicit and tacit. An interplay of the quantitative and qualitative approaches to translation / B. Lewandowska-Tymoszczyk / University of Łodz // Quantitative Methods in Corpus-Based Translation Studies: a practical guide / ed. By Michel P.Oakes and MendJi. – Amsterdam: John Benjamins Publishing, 2012. – P. 1-34.

*Стаття надійшла до редакції 16.10.2014.*

**Фокин С.Б., к.филол.н., доц.**
Институт филологии КНУ им. Т. Шевченко, Киев

## ПРЕИМУЩЕСТВА КОРПУСОВ МАЛОГО ОБЪЕМА ДЛЯ ИССЛЕДОВАНИЯ МАКРОЯВЛЕНИЙ В ПЕРЕВОДЕ

Достаточный объем корпуса может варьироваться в зависимости от целей исследования. При исследовании макроявлений в оригинальных и переведенных текстах принципиально различать явления 1) поддающиеся формализации; 2) формализируемые частично; 3) неформализируемые. Для первых и вторых целесообразно использовать корпусы малого объема, которые в данном случае представляют ряд преимуществ.

**Ключевые слова:** перевод, макростилистика, корпусная лингвистика, лингвостатистика.

**Fokin S.B., PhD, AP**
Institute of Philology, Taras Shevtchenko National University of Kyiv, Kyiv

## ADVANTAGES OF SHORT VOLUME CORPORA FOR EXPLORING MAKROPHENOMENA IN TRANSLATION

A sufficient volume of a corpus may vary according to the needs of the research. Within macrostylistic phenomena it is fundamental to differentiate 1) non-formalizable; 2) partially

formalizable; 3) formalizable ones. The first and the second group can be successfully proven on little volume corpora that can even present series of advantages in contrast to the large ones.

**Key words:** translation, macrostylistics, corpus linguistics, linguostatistics.