УДК 004.056.523:81’322:81’246.2 DOI: 10.31673/2409-7292.2026.032905

Кузавков Василь Вікторович
доктор технічних наук, професор, начальник кафедри
Військовий інститут телекомунікацій та інформатизації імені Героїв Крут, Київ, Україна
ORCID: 0000-0002-0655-9759
E-mail: vasyl.kuzavkov@viti.edu.ua

КІБЕРЛІНГВІСТИЧНА ІДЕНТИФІКАЦІЯ АКТОРА КІБЕРПРОСТОРУ НА ОСНОВІ ІНТЕРФЕРЕНЦІЙНИХ МАРКЕРІВ БІЛІНГВІЗМУ

Стаття присвячена введенню та науковому обґрунтуванню поняття кіберлінгвістичної ідентифікації як методу визначення соціолінгвістичного профілю актора кіберпростору на основі аналізу інтерференційних маркерів першої мови у його письмових текстах другою мовою. У роботі здійснено критичний аналіз наявних наукових підходів – Native Language Identification, Author Profiling та судової лінгвістики (forensic linguistics). Доведено їхню концептуальну та інструментальну недостатність для вирішення задач технічної та кіберрозвідки внаслідок відсутності цілісної математичної моделі та орієнтації переважно на якісні або вузькоприкладні освітні оцінки. Запропоновано математичну модель стану білінгвальної мовної особистості як динамічної системи. Модель формалізує згасання інтенсивності інтерференційних маркерів через систему диференціальних рівнянь першого порядку з урахуванням індивідуального коефіцієнта інтенсивності навчання, часу мовної адаптації та нормованої типологічної відстані між мовами. На основі апарату теорії інформації Шеннона, обмеженості когнітивного ресурсу робочої пам'яті та емпіричних даних обґрунтовано гіпотезу про залишкову інтерференцію: доведено, що залишкова ентропія маркерів рідної мови принципово не може дорівнювати нулю у скінченний час, що забезпечує фундаментальну розв'язність задачі ідентифікації. Сформовано векторний простір ознак тексту, сформульовано байєсівську модель пошуку соціолінгвістичного профілю та визначено межі застосовності методу як функції обсягу тексту, тривалості адаптації та спорідненості мовної пари. Практичні результати орієнтовані на інтеграцію в комплекси кіберрозвідки, аналітичні системи OSINT та засоби автоматизованої верифікації джерел інформації.

Ключові слова: кіберлінгвістична ідентифікація, актор кіберпростору, інтерференційні маркери, білінгвізм, математична модель, технічна розвідка, кіберрозвідка, лінгвістична відстань, соціолінгвістичний профіль.

Постановка проблеми

Розвиток цифрових комунікацій та інформаційного простору створює якісно нові виклики для систем технічної розвідки та кібербезпеки. Сучасні актори кіберпростору часто діють в умовах анонімності, використовують кілька мов комунікації та свідомо маскують ознаки своєї ідентичності. Традиційні методи ідентифікації, засновані на технічних артефактах (IP–адресах, метаданих, цифрових підписах), стають дедалі менш ефективними внаслідок широкого застосування засобів анонімізації.

Водночас, залишається практично недослідженим потенціал лінгвістичних характеристик письмового тексту як стійкого ідентифікаційного сигналу. Мовна особистість формується протягом тривалого часу і несе в собі відбиток рідної мови (L_1), середоваща та соціального досвіду – ознаки, які складно повністю приховати навіть при свідомому маскуванні.

Аналіз наукових публікацій свідчить про наявність розвинених окремих напрямків – *Native Language Identification* (NLI), *Author Profiling*, *forensic linguistics* [1,2,3,4]. Однак жоден з них не забезпечує комплексного підходу до задач технічної розвідки з формалізованою математичною моделлю та верифікованими межами застосовності.

Метою статті є формалізація підходу до кіберлінгвістичної ідентифікації актора кіберпростору на основі аналізу інтерференційних маркерів білінгвізму.

Для досягнення поставленої мети передбачено вирішення таких задач:

- провести аналіз існуючих підходів та виявити їхні обмеження щодо використання в інтересах технічної розвідки;

- побудувати та математично обґрунтувати модель стану білінгвальної мовної особистості як динамічної системи, що враховує закономірності адаптаційного згасання

інтерференційних маркерів першої мови, індивідуальну інтенсивність навчання та нормовану типологічну відстань між мовами;

- сформулювати та обґрунтувати гіпотезу про залишкову інтерференцію (невикорінність маркерів рідної мови) на основі апарату теорії інформації, когнітивних обмежень робочої пам'яті та емпіричних даних;

- сформувати векторний простір ознак тексту, розробити байєсівську модель пошуку соціолінгвістичного профілю актора кіберпростору та формалізувати функцію точності методу;

- визначити межі застосовності методу, окреслити напрямки практичного використання результатів у систем кіберрозвідки та OSINT.

Виклад основного матеріалу

Аналіз існуючих підходів

Огляд наукової літератури дозволяє виділити три основні напрямки досліджень у суміжних галузях.

Native Language Identification (NLI) є задачею автоматичної ідентифікації рідної мови автора на основі його текстів, написаних другою мовою. Підхід спирається на аналіз орфографічних помилок, граматичних патернів та лексичних особливостей, зумовлених впливом L_1. Змагання PAN@CLEF (з 2011 р.) та NLI Shared Tasks зафіксували точність сучасних систем на рівні 85–90% для стандартних умов. Разом з тим, NLI орієнтований переважно на освітні застосування і не враховує динаміку мовної адаптації та вплив рівня підготовки суб'єкта [1,2].

Author Profiling охоплює ширше коло задач – визначення віку, статі, рідної мови та психологічних характеристик автора за текстом. Інструменти типу LIWC (Linguistic Inquiry and Word Count) дозволяють кількісно виміряти психолінгвістичні параметри тексту. Проте існуючі підходи не забезпечують формалізованої математичної моделі об'єкту і не враховують типологічну відстань між мовами як системний параметр [3].

Судова лінгвістика (Forensic Linguistics) досліджує лінгвістичну ідентифікацію авторства у правовому контексті. Геолінгвістичне профілювання та аналіз інтерлінгвальних ознак наближаються до задач розвідки, однак залишаються переважно якісними методами без кількісної оцінки достовірності [4].

Порівняльний аналіз існуючих підходів наведено в таблиці 1. Аналіз дозволяє констатувати наявність наукової прогалини: відсутня єдина математична модель білінгвальної мовної особистості як динамічної системи, що може бути застосована для формування методу кіберлінгвістичної ідентифікації в інтересах технічної розвідки.

Таблиця 1

Порівняльний аналіз існуючих підходів

Підхід	**Переваги**	**Обмеження для задач ТР**
NLI	Висока точність для стандартних умов	Не враховує адаптацію, немає моделі об'єкту
Author Profiling	Багатовимірний профіль	Немає математичної моделі, не орієнтований на ТР
Forensic Linguistics	Практична спрямованість	Якісні методи, немає кількісної оцінки
Запропонований метод	Математична модель + паспорт точності	Потребує верифікації на розширених датасетах

Концептуальні основи кіберлінгвістичної ідентифікації

Для розбудови понятійного апарату методу запропоновано такі концептуальні визначення:

© Кузавков В.В. Кіберлінгвістична ідентифікація актора кіберпростору на основі інтерференційних маркерів білінгвізму. Сучасний захист інформації, 3(67), 68–75.
https://doi.org/10.31673/2409-7292.2026.032905

- кіберлінгвістична ідентифікація (КЛІ) – метод визначення соціолінгвістичного профілю актора кіберпростору на основі статистичного аналізу інтерференційних маркерів першої мови (L_1) у його письмових текстах, що функціонують у другій мові (L_2);
- актор кіберпростору – суб'єкт, що здійснює цілеспрямовану комунікативну діяльність у цифровому середовищі та характеризується сталими лінгвістичними параметрами, що відображають його мовну ідентичність;
- інтерференційний маркер – статистично значуще відхилення від норми L_2, зумовлене системним впливом L_1 на рівні фонетики, морфології, синтаксису або лексики;
- соціолінгвістичний профіль актора – векторна характеристика суб'єкта у просторі ознак {L_1, регіон, освіта, вік}, що визначається методом КЛІ з кількісною оцінкою достовірності.

Математична модель стану білінгвальної мовної особистості

Для формалізації процесу лінгвістичної ідентифікації білінгвальна мовна особистість розглядається як динамічна система, повний стан якої у момент часу *t* визначається вектором:

$$S(t) = \{ L_1, L_2(t), \alpha, t, D(L_1, L_2) \}, \quad (1)$$

де: L_1 – домінантність рідної мови, $L_1 \in [0, 1]$; визначається як частка мовленнєвих актів, що здійснюються рідною мовою у загальному мовленнєвому репертуарі суб'єкта; $L_2(t)$ – рівень засвоєння другої мови у момент часу *t*, $L_2 \in [0, 1]$; змінюється з часом згідно з моделлю набуття мови (4); α – індивідуальний коефіцієнт інтенсивності навчання, $\alpha > 0$; характеризує швидкість засвоєння L_2 конкретним суб'єктом (залежить від здібностей, мотивації, інтенсивності занурення); *t* – час перебування суб'єкта у мовному середовищі L_2, $t \geq 0$ [роки]; $D(L_1, L_2)$ – нормована типологічна відстань між мовами, $D \in [0, 1]$; визначається на основі лінгвістичної класифікації (підрозділ 4.3).

Вибір зазначених параметрів обґрунтований теорією засвоєння другої мови (Second Language Acquisition, SLA) [5] та контрастивною лінгвістикою [6]. Параметр α введено для обліку індивідуальних відмінностей між суб'єктами, що є принциповим для задач ідентифікації, де кожен суб'єкт розглядається як окремий об'єкт моделювання.

Вивід моделі динаміки інтерференційних маркерів

Для побудови моделі *M(t)* використано апарат звичайних диференціальних рівнянь першого порядку – стандартний інструментарій для опису процесів згасання у технічних системах.

Припущення 1 (пропорційність швидкості згасання). Миттєва швидкість зменшення інтенсивності інтерференційних маркерів *M* пропорційна їх поточній інтенсивності та залежить від індивідуальної інтенсивності навчання α:

$$dM/dt = -\gamma\, \alpha\, M(t), \quad (2)$$

де $\gamma > 0$ – коефіцієнт адаптаційного згасання, що характеризує чутливість маркерів до мовного занурення (γ є властивістю мовної пари, а не індивіда).

Припущення 1 є аналогом закону радіоактивного розпаду та закону Ньютона для охолодження, широко застосовуваних у технічних науках для опису процесів експоненційного згасання. Фізичний зміст: чим інтенсивніше та результативніше навчання (більше α), тим швидше зменшується щільність маркерів L_1.

Розв'язок. Рівняння (2) є лінійним однорідним ДР першого порядку з відокремлюваними змінними.

Розділимо змінні:

$$dM/M = -\gamma\, \alpha\, dt.$$

Інтегруємо обидві частини:

$$\int \frac{dM}{M} = -\int \gamma\, \alpha\, dt;$$

$$lnM(t) = -\gamma\, \alpha\, t + C_0,$$

де C_0 – константа інтегрування.

Потенціюємо:

$$M(t) = \hat{e}^{C_0}\, \hat{e}^{-\gamma\, \alpha\, t} = M^*\, \hat{e}^{-\gamma\, \alpha\, t}.$$

З початкової умови $M(0) = M_0$ (базова інтенсивність маркерів при $t = 0$, тобто на початку перебування у середовищі L_2) отримуємо $M^*=M_0$.

Таким чином:

$$M(t) = M_0\, \hat{e}^{-\gamma\, \alpha\, t}\ . \tag{3}$$

Масштабування на типологічну відстань. Базова інтенсивність маркерів M_0 визначається не лише індивідуальними характеристиками суб'єкта, але й структурною відстанню між мовами: для близькоспоріднених мов ($D - 0$) маркери апріорі слабкіші навіть при $t = 0$, оскільки системи L_1 і L_2 мають велику кількість спільних елементів.

Вводимо:

$$M_0 = \widehat{M_0}\, D(L_1, L_2),$$

де $\widehat{M_0}$ – нормована базова (інтенсивність, що не залежить від мовної пари).

Підставляємо у (3):

$$M(t) = \widehat{M_0}\, D(L_1, L_2)\, \hat{e}^{-\gamma\, \alpha\, t}. \tag{4}$$

Формула (4) є центральним результатом математичної моделі. Вона описує інтенсивність інтерференційних маркерів L_1 у тексті другою мовою як функцію трьох незалежних факторів: типологічної відстані між мовами D, індивідуальної інтенсивності навчання α та часу мовної адаптації t.

Фізичний зміст параметрів формули (4):

при $D \rightarrow 0$ (близькоспоріднені мови):

$M \rightarrow 0$ незалежно від t і α – маркери слабкі з самого початку;

при $D \rightarrow 1$ (неспоріднені мови):

$M = \widehat{M_0}\, \hat{e}^{-\gamma\, \alpha\, t}$ – інтенсивність визначається лише часом адаптації;

при $t \rightarrow 0$:

$M \rightarrow \widehat{M_0}\ D$ – початкова інтенсивність визначається лише типологічною відстанню;

при $t \rightarrow \infty$:

$M \rightarrow 0$ – однак, як показано у підрозділі 4.4, нуль недосяжний у скінченний час.

Вивід параметру типологічної відстані $D(L_1, L_2)$

Нормована типологічна відстань $D(L_1, L_2)$ є кількісним виразом структурної несхожості двох мовних систем. В основу її визначення покладено класифікацію мов світу за генетичною спорідненістю (мовні сім'ї та підгрупи) і типологічною структурою (морфологічний тип, синтаксичний порядок слів, фонологічний інвентар).

Для операційного визначення D можуть використовуватися відстані Левенштейна на базових словниках (ASJP-база даних) [7] або лінгвістична відстань за методологією LINGUIST List. Для цілей даної роботи пропонується нормована п'ятирівнева шкала (таблиця 2), яка відповідає загальноприйнятій генетичній класифікації.

Примітка: параметр D є властивістю мовної пари (L_1, L_2), а не суб'єкта. Він визначається один раз для кожної пари мов і залишається константою у межах конкретної задачі ідентифікації. Це принципово відрізняє D від параметрів α і t, які є індивідуальними характеристиками суб'єкта.

Таблиця 2

Параметр $D(L_1, L_2)$ для типових мовних пар

Мовна пара $L_1 \rightarrow L_2$	**Генетична група**	**$D(L_1, L_2)$**	**Інтенсивність М**
Рос. → Укр.	Слов. (сх. підгрупа)	0,10–0,15	Дуже низька
Біл. → Укр.	Слов. (сх. підгрупа)	0,10–0,20	Дуже низька
Пол. → Укр.	Слов. (різні підгрупи)	0,30–0,40	Середня
Нім. → Англ.	Германські	0,35–0,45	Середня
Франц. → Англ.	Різні гілки ІЄ	0,50–0,60	Вища за середню
Араб. → Англ.	Різні мовні сім'ї	0,85–0,95	Висока
Кит. → Англ.	Різні мовні сім'ї	0,90–1,00	Максимальна

Гіпотеза про невикорінність інтерференційних маркерів

Центральним теоретичним питанням для обґрунтування можливості КЛІ – чи може інтенсивність маркерів $M(t)$ досягти нуля при скінченному *t.*

Формально з виразу (4) випливає, що $M(t) \rightarrow 0$ лише при $t \rightarrow \infty$, тобто асимптотично. Однак це міркування спирається на математичну ідеалізацію. Для практичного обґрунтування сформулюємо гіпотезу, що спирається на теорію інформації [8].

Гіпотеза про залишкову інтерференцію виглядає наступним чином:

для будь-якої білінгвальної особистості з $L_1 \neq L_2$ залишкова ентропія маркерів першої мови у тексті другою мовою задовольняє умову:

$$H(L_1 \mid text) \geq H_{min} > 0 ,$$

де H_{min} – строго додатна нижня межа, що визначається типологічною відстанню $D(L_1, L_2)$ і є невід'ємною від когнітивної архітектури білінгва.

Обґрунтування спирається на три незалежні аргументи.

Аргумент 1 – інформаційний.

За теоремою Шеннона про кодування [9], передача інформації від джерела (мовної системи L_1) до спостерігача через канал (текст L_2) не може бути повністю пригнічена без нескінченної надлишковості коду. Будь-яке реальне кодування зберігає ненульову кількість взаємної інформації:

$$I(L_1; text) = H(L_1) - H(L_1|text) > 0 \text{ при } D > 0 . \quad (5)$$

Звідси $H(L_1; text) < H(L_1)$ (текст завжди несе ненульову інформацію про L_1).

Аргумент 2 – когнітивний.

Повна елімінація впливу L_1 на продукування тексту L_2 вимагала б нескінченного когнітивного ресурсу – постійного свідомого контролю кожного мовного рішення на всіх рівнях (фонологічному, морфологічному, синтаксичному, лексичному, прагматичному).

Когнітивний ресурс суб'єкта є скінченним (обмеженість робочої пам'яті [10]), тому частина мовних рішень неминуче приймається автоматично, під впливом домінантних патернів L_1.

Аргумент 3 – емпіричний.

Систематичний огляд понад ста публікацій з NLI [1] не виявив жодного задокументованого випадку повної елімінації маркерів L_1 у природних текстах носія L_2.

Навіть у дикторів і акторів, що пройшли тривале мовне тренування, інструментальний аналіз виявляє залишкові просодичні та синтаксичні ознаки L_1.

© Кузавков В.В. Кіберлінгвістична ідентифікація актора кіберпростору на основі інтерференційних маркерів білінгвізму. Сучасний захист інформації, 3(67), 68–75.
https://doi.org/10.31673/2409-7292.2026.032905

Таким чином, гіпотеза про залишкову інтерференцію є обґрунтованою з позицій теорії інформації, когнітивної психології та емпіричних даних. Вона є концептуальною основою, що робить задачу КЛІ принципово розв'язуваною, і разом з тим пояснює, чому точність ідентифікації є функцією, а не константою (підрозділ 5.3).

Модель набуття другої мови $L_2(t)$

Для повноти моделі (1) необхідно визначити вид функції $L_2(t)$. В теорії SLA широко використовується логістична модель [8], що описує процес набуття мови як насичення:

$$L_2(t) = L_{2max}/(1 + e^{-\alpha(t-t_0)}) , \quad (6)$$

де $L_{2max} \in (0, 1]$ – максимальний досяжний рівень засвоєння L_2 для даного суб'єкта (залежить від здібностей і умов навчання); t_0 – час досягнення половини максимального рівня (точка перегину).

З (6) випливає, що $L_2(t)$ монотонно зростає від 0 до L_{2max} при $t \to \infty$, що якісно узгоджується зі спостережуваними закономірностями засвоєння мови. Підставляючи (6) у вираз для функції точності (підрозділ 5.3), отримуємо повну залежність точності ідентифікації від часу мовної адаптації суб'єкта.

Зведена таблиця позначень і посилань прийнятих при опису математичної моделі стану білінгвальної мовної особистості представлено в таблиці 3.

Таблиця 3

Позначення та посилання

Позначення	**Зміст**	**Область**	**Джерело**
$M(t)$	інтенсивність інтерференційних маркерів L_1 у момент часу t	[0, 1]	(4)
$\widehat{M_0}$	нормована базова інтенсивність маркерів (при $t = 0, D = 1$)	[0, 1]	(4)
$D(L_1, L_2)$	нормована типологічна відстань між мовами	[0, 1]	табл. 2, [9]
γ	коефіцієнт адаптаційного згасання (властивість мовної пари)	> 0	(2)
α	індивідуальний коефіцієнт інтенсивності навчання	> 0	(1), [8]
t	час перебування у мовному середовищі L_2	роки ≥ 0	(1)
$L_2(t)$	рівень засвоєння другої мови у момент t	[0, 1]	(6), [8]
L_{2max}	максимальний досяжний рівень засвоєння L_2	(0, 1]	(6)
$H(L_1\|text)$	умовна ентропія L_1 за наявності тексту	$\geq H_{min}>0$	гіпотеза, [7]
$I(L_1; text)$	взаємна інформація між L_1 та текстом	>0 при D>0	(5), [7]

Метод кіберлінгвістичної ідентифікації

1. Ознаковий простір.

На основі математичної моделі формується вектор ознак тексту:

$$X = [x_1, x_2, \dots, x_i],$$

де компоненти вектору відображають вимірювані характеристики тексту: x_1 – частота L_1–специфічних синтаксичних патернів; x_2 – індекс лексичної складності; x_3 – коефіцієнт пунктуаційної девіації; x_4 – щільність морфологічних інтерференцій; $x_5 \dots x_i$ – додаткові ознаки відповідно до конкретної мовної пари.

2. Алгоритм ідентифікації.

Задача пошуку оптимального профілю формулюється як задача байєсівської класифікації:

$$C^* = \underset{c}{\operatorname{argmax}}\, P(C \mid x) = \underset{c}{\operatorname{argmax}} \frac{P(x \mid C) \cdot P(C)}{P(x)}, \quad (7)$$

де C – клас (соціолінгвістичний профіль); $P(C)$ – апріорна ймовірність класу (визначається до аналізу тексту, наприклад частота даної рідної мови у досліджуваній популяції); $P(x|C)$ – функція правдоподібності вектору ознак x за умови належності до класу C; $P(C|x)$ – апостеріорна ймовірність класу після спостереження вектору ознак x.

Результатом моделювання є розподіл ймовірностей по класах з оцінкою достовірності, що дозволяє відійти від точкового оцінювання.

3. Функція точності методу

Точність ідентифікації пропонується розглядати як функцію від параметрів моделі:

$$A = f(D, a, t, V), \quad (8)$$

де V – обсяг аналізованого тексту. Межі застосовності методу визначаються умовами зниження точності A нижче прийнятного порогу.

Верифікація та практичне застосування

З огляду на теоретико–методологічний характер дослідження, первинну оцінку гіпотез запропоновано здійснювати на основі відкритих джерел з відомими параметрами суб'єктів. Медійні особи (політики, журналісти, автори з міграційним досвідом) є релевантним об'єктом для аналізу за такими критеріями:

- наявність верифікованих біографічних даних (рідна мова, освіта);
- значний обсяг доступного текстового корпусу в цифровому середовищі;
- жанрове різноманіття текстів.

Практичне значення отриманих результатів полягає у можливості інтеграції запропонованого підходу в системи OSINT, комплекси кіберрозвідки, інформаційно–аналітичні системи та засоби автоматизованого оцінювання соціолінгвістичних характеристик суб’єктів цифрової комунікації [11].

Обмеження дослідження

При розробці та застосуванні моделі слід враховувати такі обмеження:

- залежність від обсягу даних **(**точність ідентифікації суттєво залежить від обсягу (V) доступного текстового корпусу);
- складність мовних пар (виявлення маркерів значно ускладнюється у випадку близькоспоріднених мов (наприклад, українська–російська), де параметр D прямує до мінімуму);
- навмисне маскування (застосування суб'єктом спеціалізованих інструментів чи свідомого лінгвістичного маскування може спотворювати щільність маркерів);
- потреба емпіричних даних (модель потребує подальшої розширеної емпіричної верифікації на диверсифікованих датасетах для уточнення вагових коефіцієнтів).

Висновки

У роботі запропоновано концептуальний підхід до кіберлінгвістичної ідентифікації в інтересах інформаційно–технічних систем. Основними теоретичними результатами є:

- побудова математичної моделі білінгвальної мовної особистості як динамічної системи, що враховує параметри навчання та лінгвістичної відстані;

- формалізація принципу залишкової інтерференції через оцінку ентропійних характеристик тексту.

Введення параметру типологічної відстані $D(L_1, L_2)$ дозволяє диференціювати потенційну точність аналізу для різних мовних пар. Подальші напрямки досліджень пов'язані з калібруванням коефіцієнтів моделі на розширених корпусах текстів.

Перелік посилань

1. Goswami D., Thilagan S., North K., Malmasi S., Zampieri M. 2024. Native Language Identification in Texts: A Survey. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational

© Кузавков В.В. Кіберлінгвістична ідентифікація актора кіберпростору на основі інтерференційних маркерів білінгвізму. Сучасний захист інформації, 3(67), 68–75.
https://doi.org/10.31673/2409-7292.2026.032905

Linguistics: Human Language Technologies (Vol. 1: Long Papers), P. 3149-3160, Mexico City, Mexico. Association for Computational Linguistics. URL: https://aclanthology.org/2024.naacl-long.173/ https://doi.org/10.18653/v1/2024.naacl-long.173.

2. Markov I, Nastase V, Strapparava C. Exploiting native language interference for native language identification. Natural Language Engineering. 2022. Vol. 28 no. 2. P. 167-197. URL: https://doi.org/10.1017/S1351324920000595.

3. Bevendorff, J. et al. (2023). Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection: Condensed Lab Overview. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2023). Lecture Notes in Computer Science, vol 14187. Springer, Cham. https://doi.org/10.1007/978-3-031-42448-9_29.

4. Roemling D., Scherrer Y., Miletić A. Explainability of machine learning approaches in forensic linguistics: a case study in geolinguistic authorship profiling. Proceedings of the 1st Workshop on NLP for AI, Cyber Security and Forensic Linguistics (NLP4AICS 2024). Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 1–12. URL: https://doi.org/10.18653/v1/2024.nlp4aics-1.1.

5. Krashen S.D. Principles and Practice in Second Language Acquisition. Oxford: Pergamon Press, 1982. 202 p.

6. Jarvis S., Pavlenko A. Crosslinguistic influence in language and cognition. New York: Routledge, 2008. 264 p. URL: https://doi.org/10.4324/9780203935927

7. Wichmann S. et al. The ASJP Database (v19). 2020. URL: https://asjp.clld.org

8. Cover T.M., Thomas J.A. Elements of Information Theory. 2nd ed. Wiley, 2006. 748 p.

9. Shannon C.E. A Mathematical Theory of Communication // Bell System Technical Journal. 1948. – Vol. 27. P. 379–423.

10. Baddeley A. Working Memory, Thought, and Action. Oxford University Press, 2007. 412 p.

11. Williams H., Blum I. Defining Second Generation Open-Source Intelligence (OSINT) for the Defense Enterprise. – RAND Corporation, 2018. URL: https://doi.org/10.7249/RR1964

Vasyl Kuzavkov
Doctor of Technical Sciences, Professor, Head of Department
Military Institute of Telecommunications and Information Technologies named after the Heroes of Kruty, Kyiv, Ukraine
ORCID: 0000-0002-0655-9759
E-mail: vasyl.kuzavkov@viti.edu.ua

CYBERLINGUISTIC IDENTIFICATION OF A CYBERSPACE ACTOR BASED ON INTERFERENCE MARKS OF BILINGUALISM

The article is devoted to the introduction and scientific substantiation of the concept of cyberlinguistic identification as a method for determining the sociolinguistic profile of a cyberspace actor based on the analysis of interference markers of their first language in written texts composed in a second language. A critical analysis of existing scientific approaches – Native Language Identification, Author Profiling, and forensic linguistics – is performed. Their conceptual and instrumental insufficiency for solving technical intelligence and cyber intelligence tasks is demonstrated due to the lack of a holistic mathematical model and a primary focus on qualitative assessments or narrow educational applications. A mathematical model of the state of a bilingual linguistic personality as a dynamic system is proposed. The model formalizes the decay of interference marker intensity using a system of first-order differential equations, taking into account the individual learning rate coefficient, language adaptation time, and normalized typological distance between languages. Based on Shannon's information theory framework, the cognitive limitations of working memory capacity, and empirical data, the residual interference hypothesis is substantiated: it is proven that the residual entropy of native language markers fundamentally cannot equal zero in finite time, which provides the fundamental solvability of the identification problem. A text feature vector space is formed, a Bayesian model for searching the sociolinguistic profile is formulated, and the applicability boundaries of the method are defined as a function of text volume, duration of adaptation, and language pair kinship. Practical results are oriented toward integration into cyber intelligence suites, OSINT analytical systems, and tools for automated verification of information sources.

Keywords: cyberlinguistic identification, cyberspace actor, interference markers, bilingualism, mathematical model, technical intelligence, cyber intelligence, linguistic distance, sociolinguistic profile.

Надійшла до редакції (Received): 20.07.2026
Прийнята до друку (Accepted): 08.09.2026
Опубліковано онлайн (Available online): 15.09.2026

© Кузавков В.В. Кіберлінгвістична ідентифікація актора кіберпростору на основі інтерференційних маркерів білінгвізму. Сучасний захист інформації, 3(67), 68–75.
https://doi.org/10.31673/2409-7292.2026.032905