

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА
АКАДЕМІЯ»

Кафедра інформатики факультету інформатики



**МЕТОДИ ВВЕДЕННЯ (РЕДАКТОР ІМЕ) ДЛЯ ПІДТРИМКИ
БАГАТОМОВНОСТІ ВЕБ-САЙТІВ**

**Текстова частина до курсової роботи
за спеціальністю «Комп'ютерні науки» 122**

Керівник курсової роботи
д.т.н, доц. Глибовець А. М.

_____ (підпис)
«____» _____ 2020 р.

Виконав студент 3 р.н. БП
«Комп'ютерні науки»
Галайда М. В.

«____» _____ 2021 р.

Міністерство освіти і науки України

Київ 2021

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА
АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Керівник курсової роботи
д.т.н, доц. Глибовець А. М.

(підпис)

«____» _____ 2020 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студенту 3-го курсу, факультету інформатики
Галайді Михайлу Вікторовичу

ТЕМА Методи введення (редактор ІМЕ) для підтримки
багатомовності веб-сайтів

Зміст ТЧ до курсової роботи:

Зміст

Анотація

Вступ

1 Лінгвістика та особливості світових мов

2 Зберігання цифрової інформації різними мовами

3 Введення цифрових даних для різних систем письма

4 Реалізація та огляд власного ІМЕ

Висновки

Список літератури

Додатки

Дата видачі «____» _____ 2020 р.

Керівник _____ (підпис)

Завдання отримав _____ (підпис)

Тема: розробка редактора методу введення

Календарний план роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1	Отримання завдання на курсову роботу	15.10.2020	
2	Огляд літератури за темою роботи	10.02.2021	
3	Дослідження лінгвістичних особливостей мов	18.03.2021	
4	Аналіз сучасних ІМЕ	23.03.2021	
5	Реалізація власного ІМЕ	29.03.2021	
6	Написання пояснювальної роботи	01.04.2021	
7	Написання доповіді та презентації	05.04.2021	
8	Аналіз результатів із науковим керівником	13.04.2021	
9	Коригування роботи за результатами попереднього огляду	14.04.2021	
10	Остаточне оформлення пояснювальної роботи та презентації	15.05.2021	
11	Захист курсової роботи	17.05.2021	

Студент Галайда М. В.

Керівник Глибовець А. М.

ЗМІСТ

Анотація.....	6
Вступ	7
РОЗДІЛ 1: Лінгвістика та особливості світових мов.....	10
1.1. Розуміння контексту мови для побудови ІМЕ	10
1.2. Обґрунтування вибору мов для дослідження.....	11
1.3. Лінгвістичний огляд обраних мов	12
1.3.1.1. Мандаринська мова	12
1.3.2. Японська мова	17
1.3.3. Арабська мова	20
1.3.4. В'єтнамська мова	22
1.4. Висновки за розділом 1.....	24
РОЗДІЛ 2: Зберігання цифрової інформації різними мовами	25
2.1. Кодування символів у цифрових системах.....	25
2.2. Композиція складних символів у системах кодувань	27
2.3. Висновки за розділом 2.....	28
РОЗДІЛ 3: Введення цифрових даних для різних систем письма	29
3.1. Опис поняття та принципів роботи ІМЕ.....	29
3.1.1. Фонетичний метод введення	30
3.1.2. Огляд та використання методу на прикладі досліджуваних мов	30
3.1.3. Реалізація методу в різних ІМЕ.....	31
3.2. Методи введення, що опираються на розкладку клавіатури	33
3.2.1. Огляд та використання методів на прикладі досліджуваних мов	33
3.2.2. Реалізація методу в різних ІМЕ.....	35
3.3. Інші методи введення	36
3.3.1. Огляд та використання методів на прикладі досліджуваних мов	36

3.3.2. Реалізація методів у різних ІМЕ	38
3.4. Висновки за розділом 3.....	39
РОЗДІЛ 4: Реалізація та огляд власного ІМЕ.....	40
4.1. Опис та обґрунтування формату запропонованого рішення	40
4.2. Технології та інструменти, обрані для реалізації.....	40
4.3. Архітектура та реалізація застосунку	41
4.4. Демонстрація роботи застосунку.....	44
4.5. Висновки за розділом 4.....	46
Висновки.....	47
Література.....	48
Додаток А (довідниковий) Блоки Юнікоду, повністю відведені під комбінувальні символи	49
Додаток Б (довідниковий) Графічні приклади реалізації методів введення в різних ІМЕ	51

Анотація

Метою курсової роботи є дослідити проблематику цифрового введення й збереження інформації різними світовими мовами.

Було досліджено лінгвістичні особливості декількох мов та розглянуто принципи введення та збереження символів цих мов.

Проаналізовано сучасні системи кодування символів, їх особливості та переваги. Вивчено принципи комбінування початкових символів з метою створення нових.

Досліджено принципи роботи редакторів методів введення та проаналізовано наявні приклади.

З урахуванням результатів дослідження було розроблено власний редактор методу введення для полегшеного введення складених символів у веб-браузері.

Ключові слова: редактор методу введення, кодування символів, комбінування символів.

Вступ

Швидке поширення комп'ютерів та Інтернету вимагає зручних та ефективних методів взаємодії незалежно від мови користувача. В той же час, персональні комп'ютери у вигляді, у якому ми їх знаємо сьогодні, розроблялися для мов європейського походження: наприклад, одна із найпоширеніших розкладок клавіатури QWERTY, яку вперше використовували ще у 19-му сторіччі, адаптована саме для англійської та інших європейських мов.

Такі підходи, як стало зрозуміло пізніше, не були прийнятними для принципово інших за будовою мов, таких як: китайська (група різновидів мови), японська, мови індійського субконтиненту, арабська, тощо. При цьому, велика частка користувачів всесвітньої мережі походить саме із таких країн, для яких європейські принципи введення не є адаптованими. Наприклад, станом на 2021 рік, у Китайській Народній Республіці нараховується більше 900 млн активних користувачів, у Індії – більше 750 млн, 118 млн – у Японії, 112 млн та 100 млн у Бангладеш та Пакистані відповідно. Лише із перелічених вище країн нараховується близько 2 млрд користувачів, для яких не є зручними звичні для Західного світу методи введення – їхні мови часто використовують абсолютно інші підходи до побудови символів, синтаксису, семантики речень тощо.

З огляду на таку ситуацію, з'явилася гостра необхідність створення актуальних для решти світу методів введення цифрової інформації. Почали з'являтися адаптовані методи для неєвропейських мов, наприклад:

- **Пін'їнь** – метод введення китайської мови (*мандаринського діалекту*) з допомогою латинізації, розроблений у 50-их роках ХХ століття та затверджений до широкого використання у Китайській Народній Республіці у 1979-му році.

- **Ромаджі** – подібний за принципом до піньїню метод введення японської мови, що також використовує романізацію (*латинізацію*) слів та символів. Перші спроби романізації японської мови були здійснені ще у XVI столітті. Найпоширеніший сьогодні стандарт ромаджі був створений у 1886 році американським лікарем та місіонером Джеймсом Кертісом Гепберном.
- **Телекс** – один із найпоширеніших методів введення в'єтнамської мови, багатої на різноманітні діакритичні знаки, що у різних комбінаціях із базовими символами алфавіту дозволяють будувати більше сотні додаткових складених символів. Телекс набув популярності у другій половині XX століття. В основі методу лежить побудова складних символів через введення базового символу та комбінації певних клавіш клавіатури для додавання потрібних діакритиків.

Наведені вище методи введення є лише одними із багатьох – за декілька останніх десятиліть було створено сотні різноманітних методів введення для десятків світових мов. Згодом, такі методи введення почали групувати у програмні додатки, що стали називатися редакторами методів введення (Input Method Editor, IME).

За *мету* даної курсової роботи було поставлено створення актуального для поточного ринку редактору методів введення, що використовуватиме найефективніші підходи галузі.

Відповідно до заданої мети було визначено такі *завдання*:

1. Розглянути лінгвістичні особливості декількох обраних для дослідження мов.
2. Дослідити наявні для визначених мов принципи збереження та композиції символів.

3. Проаналізувати сучасні редактори методів введення, визначити переваги та недоліки в їхній реалізації.
4. З урахуванням результатів дослідження створити редактор методів введення.

Робота складається з чотирьох розділів.

У першому розділі обґрунтовується необхідність розуміння принципів мов та їх письма для якісного аналізу сучасних редакторів методів введення та подальшої розробки власного. Для дослідження обрано 4 мови: мандаринська, японська, арабська та в'єтнамська. Для кожної з мов розглянуто особливості її письма, фонетики, алфавіту, тощо.

У другому розділі проводиться огляд сучасних систем кодування символів, їх ефективності та особливості роботи зі складеними символами.

У третьому розділі розглядаються принципи роботи редакторів методів введення (IME) та окремих типів методів введення. Для кожного з типів наводяться приклади реалізації в окремих мовах й в конкретних редакторах.

У четвертому розділі знання, набуті в попередніх трьох, використовуються для розробки власного редактору методів введення, проводиться детальний огляд його алгоритмів та демонструється робота застосунку на практиці.

РОЗДІЛ 1: Лінгвістика та особливості світових мов

1.1. Розуміння контексту мови для побудови ІМЕ

Розробка будь-якого якісного програмного застосунку повинна починатися із детального дослідження предметної області та чіткого розуміння проблеми, яку застосунок покликаний вирішувати.

У випадку даної роботи, для оцінки уже наявних редакторів методів введення та побудови власного важливим є розуміти, як працюють досліджувані мови. Варто знати відповіді як на базові питання, такі як:

- Чи є письмо мови логографічним, чи ідеографічним?
- Який напрямок письма використовується (зліва направо, справа наліво, зверху вниз)?

Так і на питання, що пояснюють більш поглиблені аспекти мови, як-от:

- Як та в якій послідовності відбувається побудова символів?
- Чи міняється форма символу в залежності від його розташування у слові чи реченні?
- Чи використовуються на письмі явні голосні звуки?

Маючи відповіді на наведені вище питання, розробники ІМЕ здатні набагато ефективніше проаналізувати доступні на ринку аналоги та запропонувати своє рішення, що буде максимально пристосованим до досліджуваної мови та надаватиме якнайкращий результат у використанні.

Тому, поточний розділ присвячений саме лінгвістичному дослідженню кількох мов, вибір (та його обґрунтування) яких описаний у наступному розділі.

1.2. Обґрунтування вибору мов для дослідження

Аби найкраще зрозуміти проблеми, пов'язані із введенням неєвропейських мов, в межах даної курсової роботи було поставлено завдання дослідити мови, які:

- Використовуються великою часткою населення Землі
- Не належать до германських чи романських мов

Тому, для дослідження було обрано такі 4 мови:

1. **Мандаринська** – група китайських діалектів, що походить з Північного Китаю й історично використовувалася у Північному та Південно-Західному Китаї. Наразі є найбільш використовуваною мовою у світі, визнана офіційною мовою у КНР, Тайвані та Сингапурі й є однією із шести офіційних мов ООН. Станом на 2017-ий рік, Мандаринську мову вважають рідною понад 70% громадян КНР, що перевищує 900 млн. осіб, а іще для 200 млн. людей мандаринська є другою мовою. Загалом, частка носіїв мандаринської мови становить більше 14% усього населення Землі [1]. Настільки висока кількість носіїв однозначно вказує на важливість створення ефективних методів введення, а ключові відмінності від європейських мов роблять мандаринську мову цікавою для дослідження у даній роботі.

2. **Японська** – мова із японо-рюкюської групи, що є рідною для жителів практично усієї Японії. Певний період свого розвитку перебувала під впливом китайських мов, що врешті відбилося на лексиці та письмі. Станом на 2020-ий рік, близько 126 млн. осіб вважають японську мову рідною, а ще 1 млн. – другою [2]. Разом з великим числом носіїв, на важливість дослідження японської також впливає роль Японії у технологічному розвитку людства протягом останніх декількох десятиліть.

3. **Арабська** – семітська мова, ранні докази існування якої датуються від VI ст. до н. е. до IV ст. н. е. Арабська мова використовується у всьому Арабському світі, визнана офіційною у 26 державах та є однією із шести офіційних мов ООН. Станом на 2021-й рік, нараховується приблизно 422 млн. носіїв арабської мови та її відгалужень [3]. В той же час, арабська мова є мовою богослужіння для близько 1.8 млрд. мусульман зі всього світу. Відмінність арабської мови від своїх європейських сусідів та її широке використання роблять її очевидним кандидатом для дослідження.

4. **В'єтнамська** – найбільш вживана австроазійська мова, що виникла та активно використовується у В'єтнамі. В'єтнамською також активно розмовляє діаспора у решті Азії, Європі, Північній Америці та Австралії. На сьогодні, рідною в'єтнамську вважають близько 80 млн. носіїв [1]. Історично, в'єтнамська мова використовувала запозичені з китайських мов символи, але, внаслідок французької колонізації, зрештою перейшла на нову систему письма, що базується на латинському алфавіті.

1.3. Лінгвістичний огляд обраних мов

1.3.1. Мандаринська мова

Під **Мандаринською мовою** зазвичай мається на увазі найбільша діалектична група китайських мов (таких груп нараховується від 7 до 13, в залежності від джерела). У склад групи зокрема входить Пекінський діалект, що є основою для фонології Стандартної Китайської мови, Путунхуа – вона наразі вважається *lingua franca*, тобто мовою міжетнічної комунікації, для носіїв різних варіацій мандаринської та інших мов Китаю (Південноміньська, Кантонська, тощо).

Як і решта китайських мов, мандаринська використовує **логографічну систему** китайського письма: на відміну від фонографічного

письма, де кожен символ позначає фонетичний чи фонологічний елемент без прив'язки до будь-якого значення, у логографічному письмі окремі символи позначають цілі слова та концепти (рисунки 1.1). Наприклад, сполучення символів 漢字 (пін'їнь: hàn zì) означає «китайський символ». Окремо ж, символ 漢 означає «Китай», а 字 позначає «символ» [4].

#	Strokes	Pinyin	Meaning
1	一	yī	one
2	丨	gǔn	line
3	丶	zhǔ	stroke
4	ㄣ	fú	oblique
5	乙	yǐ	sickle
6	乚	jué	hook
7	二	èr	two
8	厶	tóu	shelter, head
9	人	rén	man
10	儿	er	child
11	入	rù	to enter
12	八	bā	eight, separate
13	口	jiǒng	scope
14	冂	mì	to cover
15	冫	bīng	ice
16	几	jǐ	table
17	匚	qiān	container
18	刀	dāo	knife
19	力	lì	force
20	勹	bāo	to wrap
21	匕	bǐ	spoon, man overthrown
22	匚	fāng	open box
23	匚	xī	to hide
24	十	shí	ten
25	卜	bo	divination

#	Strokes	Pinyin	Meaning
26	卩	jié	seal
27	厂	chǎng	production facility
28	厶	sī	secret
29	又	yòu	still, hand
30	口	kǒu	mouth
31	口	wéi	enclosure
32	土	tǔ	earth
33	士	shì	scholar
34	夕	zhī	to go
35	夕	suī	go slowly
36	夕	xī	evening
37	大	dà	big
38	女	nǚ	woman
39	子	zǐ	child
40	宀	mián	roof
41	寸	cùn	thumb
42	小	xiǎo	small
43	尢	yǒu	weak
44	尸	shǐ	dead body
45	屮	chè	germ
46	山	shān	mountain
47	川	chuān	river
48	工	gōng	work
49	己	jǐ	personal
50	巾	jīn	turban

#	Strokes	Pinyin	Meaning
51	干	gān	dry
52	么	yǎo	tiny
53	广	guǎng	shelter
54	彳	yín	great, stride
55	升	gōng	two, hands
56	弋	yì	arrow-grapin
57	弓	gōng	bow
58	彡	jì	snout
59	彡	shān	beard, brush
60	彳	chì	walk around
61	心	xīn	heart
62	戈	gē	halberd
63	户	hù	gate
64	手	shǒu	hand
65	支	zhī	branch
66	支	pū	to bump, hand
67	文	wén	writing
68	斗	dǒu	measurer
69	斤	jīn	ax, 500 grams
70	方	fāng	square, direction
71	无	wú	without
72	日	rì	sun
73	日	yuē	to say
74	月	yuè	moon
75	木	mù	tree

Рисунок 1.1 - приклади китайських логограм

Оскільки письмо є логографічним і вимова символів жодним чином не пояснює їх написання, кожен символ повинен вивчатися носієм окремо.

Незважаючи на те, що кожен символ китайського письма є логографічним, можна виокремити 6 окремих типів символів, залежно від принципу їх побудови [4]:

1. **Піктограми** (象形, пін'їнь: xiàng xíng) – символи, що графічно зображають об'єкт, який позначають. Наприклад: 人 – «людина», 日 – «сонце», 木 – «дерево».
2. **Ідеографи** або **індикативи** (指事, пін'їнь: zhǐ shì) – символи позначають абстрактні концепти, як-от: 上 – «догори», 下 – «донизу», 三 – «три».

3. **Логічні агрегати** (會意/会意, піньїнь: huìyì) – сполучення, в яких два чи більше символи використовуються для надання одного значення. Наприклад: 東 / 东 означає «схід», а самі символи зображають сонце, що сходить над деревами.
4. **Фонетичні комплекси** (形聲/形声, піньїнь: xíngshēng) складаються з радикалів, що позначають загальну семантичну категорію символу (наприклад, пов'язаність з водою, землею, тощо), та інших символів, що несуть фонетичну цінність для слова. 晴 означає «ясний» (про погоду), і складається з таких символів: 日 («сонце») та 青 («синій/зелений») – останній впливає на вимову слова.
5. **Трансференція, передача** (轉注/转注, піньїнь: zhuǎnzhù) – деякі символи, що початково позначали прості та чітко визначені поняття, поступово набули більш абстрактних та широких значень. Для прикладу, символ 網 / 网 означає «сітка» та був утворений як піктограма рибацької сітки. З часом, цей символ почав використовуватися для позначень інших сіток/мереж, як наприклад у значенні «комп'ютерна мережа».
6. **Запозичення** (假借, піньїнь: jiǎjiè) – використання символу у інакшому від початково задуманого сенсі. 哥 спершу означав «спів», «співати», однак зараз використовується для позначення «брат». Наразі ж для позначення «співу» використовується інший символ: 歌.

Нарешті, символи китайського письма можна розбити у базові одиниці – **риски**. Такі риски поділяються на вісім категорій в залежності від напрямку та форми (таблиця 1.1).

Символ	Китайська назва	Опис
一	Héng, 横	Риска праворуч
丨	Shù, 竖	Риска униз
丿	Wān, 弯	Риска, загнута ліворуч
㇏	Nà, 捺	Риска, спадна вправо
㇏	Tí, 提	Зростаюча риска
丶	Diǎn, 点	Крапка
㇏	Gōu, 钩	Гачок, загострює інші риси
-	Zhé, 折	Зміна напрямку риси

Таблиця 1.1 – типи рисок у китайських символах

Для використання таких рисок при побудові символів існують такі правила (деякі можуть оминатися залежно від ситуації):

1. Горизонтальні риси пишуться перед вертикальними, а загнуті ліворуч – перед спадними праворуч.
2. Символи будуються зверху вниз, зліва направо.
3. У симетричних символах спершу будується середина, вже потім – краї.

Історично, написання слів та речень китайським письмом відбувалося зверху вниз, справа наліво (перша колонка розташована по правий бік сторінки). Однак, оскільки китайські символи не прив'язані один до одного, вони не потребують чітких обмежень у **напрямку письма**. Зокрема тому з часом почало також використовуватися більш звичне для європейських мов письмо зліва направо, зверху вниз. Більше того, у 1955-му році влада КНР зробила такий варіант написання офіційним, а у 2004-му те саме зробила

влада Тайваню. Тим не менше, вертикальне письмо все ще використовується там, де це необхідно.

Усі китайські мови є тональними – залежно від тону, з яким вимовляється слово, його сенс може бути абсолютно різним. Наприклад, слово з вимовою та залежно від тону може означати «матір», «кінь» або «сварити». Наприклад, мандаринська група використовує 4 основні тони та 1 нейтральний: високий, зростаючий, падаючий-зростаючий, падаючий та нейтральний тони (рисунк 1.2).

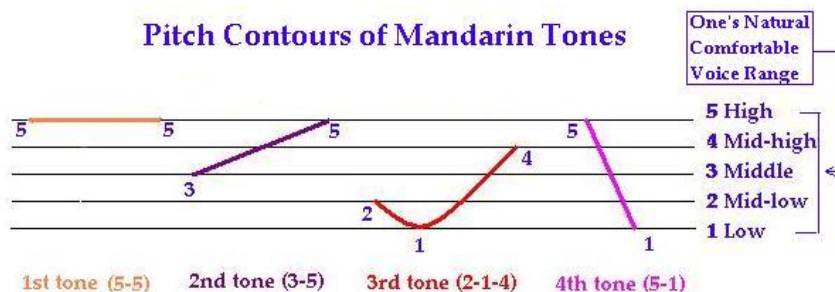


Рисунок 1.2 - візуалізація тонів мандаринської мови

У кантонській мові, наприклад, використовується 6 різних тонів: високий рівний, середній зростаючий, середній рівний, низький спадаючий, низький зростаючий та низький рівний (рисунк 1.3).

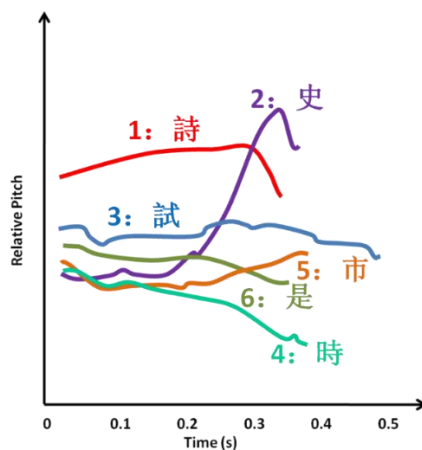


Рисунок 1.3 - візуалізація тонів кантонської мови

Важливою особливістю китайської системи письма є наявність двох варіантів написання символів: **традиційні** та **спрощені** (рисунок 1.4). Спрощені символи були введені владою КНР у 50-их та 60-их роках XX століття з метою підвищення рівня освіченості населення – такі символи вважаються простішими для вивчення, хоч для освічених носіїв мови написання традиційних символів зазвичай не є проблемою. Наразі спрощена система письма активно використовується в КНР, Сингапурі та китайській діаспорі, а традиційним символам продовжують надавати перевагу мешканці Тайваню, Гонконгу та Макау [4].

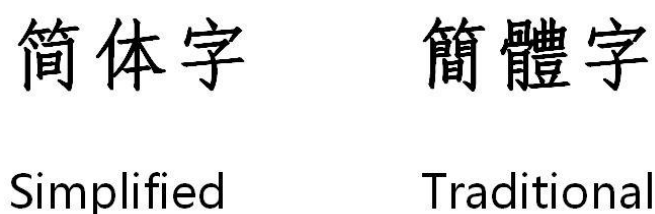


Рисунок 1.4 - порівняння традиційних та спрощених китайських символів

1.3.2. Японська мова

Для розуміння того, як побудована **японська мова**, варто спершу знати її історію. Дуже мало відомо про японську мову у час її раннього розвитку, зокрема у період Яйой – епоху історії Японії, що науковці зазвичай датують від 300 р. до н. е. до 300 р. н. е.. Відсутність будь-яких записів японською мовою з тих часів великою мірою пояснюється тим, що вона не була писемною.

У III столітті Японія встановила перші стосунки з Китаєм та знаходилася під його потужним культурним впливом протягом 600 наступних років. Враховуючи, що на той момент японська мова не мала

власної системи письма, китайська поступово стала літературною мовою Японії. Зрештою її символи почали адаптовувати та використовувати для написання японської. У VIII – IX століттях на основі адаптованих китайських символів почали утворюватися дві спрощені системи фонетичних символів **Кана** (яп. 仮名): **Катака́на** (яп. 片仮名, かたかな) та **Хірага́на** (яп. 平仮名, ひらがな). Обидві з утворених систем містять в собі 46 базових символів, та загалом можуть утворювати 71 символ з допомогою діакритичних знаків (*рисунок 1.5*). При цьому, японська мова не відмовилася від запозиченої з Китаю системи письма, і тому досі використовує декілька тисяч символів китайського походження, що об'єднані в систему **Канджі** (яп. 漢字). Таким чином, у наш час письмо японської мови є водночас логографічним (китайські символи канджі) та фонографічним (символи кана) [5].

У японській мові система канджі зазвичай використовується для написання багатьох слів китайського та японського походження, а саме: більшість іменників, основи більшості дієслів, прикметників та прислівників й переважна більшість власних японських назв та імен.

Натомість хірагана здебільшого призначена для граматичних елементів речень (словозмін, часток), уточнення неоднозначних символів канджі або й для повноцінної заміни деяких складних канджі.

Катакана ж переважно використовується для транслітерації запозичених слів некитайського походження. Також може використовуватися для позначення загальних назв тварин та рослин, звуконаслідування та акцентування окремих слів.

Hiragana

n	w-	r-	y-	m-	h-	n-	t-	s-	k-	
ん	わ	ら	や	ま	は	な	だ	さ	か	あ
N	WA	RA	YA	MA	HA	NA	TA	SA	KA	A
	ゐ	り		み	ひ	に	ち	し	き	い
	WI	RI		MI	HI	NI	CHI	SHI	KI	I
		る	ゆ	む	ふ	ぬ	つ	ず	く	う
		RU	YU	MU	FU	NU	TSU	SU	KU	U
	ゑ	れ		め	へ	ね	て	せ	け	え
	WE	RE		ME	HE	NE	TE	SE	KE	E
	を	ろ	よ	も	ほ	の	と	そ	こ	お
	WO	RO	YO	MO	HO	NO	TO	SO	KO	O

Katakana

n	w-	r-	y-	m-	h-	n-	t-	s-	k-	
ン	ワ	ラ	ヤ	マ	ハ	ナ	タ	サ	カ	ア
N	WA	RA	YA	MA	HA	NA	TA	SA	KA	A
	ヰ	リ		ミ	ヒ	ニ	チ	シ	キ	イ
	WI	RI		MI	HI	NI	CHI	SHI	KI	I
		ル	ユ	ム	フ	ヌ	ツ	ス	ク	ウ
		RU	YU	MU	FU	NU	TSU	SU	KU	U
	ヱ	レ		メ	ヘ	ネ	テ	セ	ケ	エ
	WE	RE		ME	HE	NE	TE	SE	KE	E
	ヲ	ロ	ヨ	モ	ホ	ノ	ト	ソ	コ	オ
	WO	RO	YO	MO	HO	NO	TO	SO	KO	O

Рисунок 1.5 - японські системи письма хірагана та катакана

Оскільки всі три системи письма можуть використовуватися для різних частин речення, типове речення у японській мові зазвичай поєднує в собі символи канджі та хірагана, а часто й катакана (рисунк 1.6).



Рисунок 1.6 - поєднання трьох систем письма в одному реченні японською мовою

Традиційно, у японській мові використовується форма написання татегакі (яп. 縦書き), у якій символи пишуться вертикально, у стовпчики, впорядковані зправа наліво – така форма була запозичена із традиційної китайської системи письма. Однак, сучасна японська мова також використовує форму йокогакі (яп. 横書き), у якій символи пишуться горизонтально, зліва направо [5].

У японській мові існує всього 5 голосних звуків: а, і, у, е, о – всі п'ять голосних також мають довгі форми: ā, ī, ū, ē, ō. Відмінність між довгою та короткою формами голосних є важливою, оскільки вона впливає на значення слів. Ці голосні завжди вимовляються повністю та не скорочуються, як це буває в інших мовах, наприклад, в англійській. Склади у японській мові завжди (з кількома винятками) закінчуються на голосні звуки, що мінімізує збіги приголосних та часто полегшує вимову [5].

Не менш важливою особливістю японської мови є музичний наголос. Японські мори (ритмічні одиниці, з яких складаються катакана та хірагана) можуть набувати високого або низького тону – високий тон позначає акцент у слові. Різний наголос також може сильно впливати на значення слова.

1.3.3. Арабська мова

Існує багато різновидів **арабської мови**, здебільшого обумовлених територіальними відмінностями. Тим не менше, серед них можна чітко виокремити класичну арабську мову, на якій був написаний Коран – саме з релігійних міркувань багато вірян вважають цей різновид арабської найдосконалішим.

Поширеною також є сучасна стандартна арабська мова – форма арабської, що сьогодні використовується як офіційна мова та базується на літературній мові, що пішла саме від класичної арабської. Від своїх попередників сучасна арабська великою мірою відрізняється адаптованою та розширеною лексикою, що дозволяє використовувати її як мову літератури, освіти, медіа, тощо.

Письмо арабської мови є фонографічним – кожен символ у алфавіті позначає певний приголосний звук, короткі голосні на письмі не використовуються (або для їх позначення використовуються спеціальні

діакритики – ḥarakāt, حَرَكَات), а довгі голосні та дифтонги можуть позначатися неоднозначно (рисунок 1.7) [3].

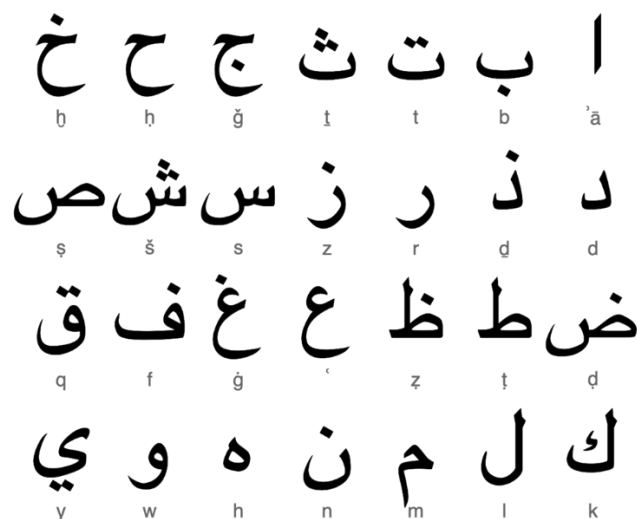


Рисунок 1.7 - арабський алфавіт

Написання арабської відбувається справа наліво. Арабські літери імітують рукописне письмо, і тому більшість із них з'єднані із сусідніми літерами, хоч деякі і залишаються ізольованими. Такий підхід пояснює наявність у арабському письмі 2-ох різних форм написання символів: **коротка** (символ на початку чи посередині слова) та **довга** (символ на кінці слова, або ізольований) (рисунок 1.8) [3].

letter	alone	beginning	middle	end
ba	ب	بـ	بـ	بـ
lam	ل	لـ	لـ	لـ
dal	د	دـ	دـ	دـ
ya	ي	يـ	يـ	يـ

Рисунок 1.8 - форми символів арабської мови в залежності від позиції у реченні

1.3.4. В'єтнамська мова

На перший погляд, сучасна **в'єтнамська мова** може здатися простішою, ніж мандаринська, японська чи арабська. Авжеж, в'єтнамський алфавіт (в'єт. Chữ Quốc Ngữ) побудований на латиниці й складається із 29 літер: 17 позначають приголосні звуки і 12 – голосні (рисунок 1.9). Серед звичних для латиниці літер, як-от *a, b, c*, тощо, є також 7 видозмінених: *ă, â, ã, đ, ê, ô, ơ*, та *ư* (рисунок 1.10) [1].

Aa	Ăă	Ââ	Bb	Cc
Dd	Đđ	Ee	Êê	Gg
Hh	Ii	Kk	Ll	Mm
Nn	Oo	Ôô	Ơơ	Pp
Qq	Rr	Ss	Tt	Uu
Ưư	Vv	Xx	Yy	

Рисунок 1.9 - в'єтнамська абетка

Áá	Àà	Ảả	Ãã	Ạạ	Òò	Ỏỏ	Õõ	Ọọ	Ôô
Ăă	Ằằ	Ẵẵ	Ẳẳ	Ằằ	Ốố	Ồồ	Ổổ	Ỗỗ	Ộộ
Ằằ	Ằằ	Ằằ	Ằằ	Ằằ	Ơơ	Ớớ	Ờờ	Ởở	Ỡỡ
Ằằ	Ằằ	Đđ	Éé	Èè	Ợợ	Úú	Ùù	Ủủ	Ûủ
Ằằ	Ằằ	Ằằ	Êê	Ếế	Ựự	Ưư	Ứứ	Ừừ	Ửử
Ằằ	Ằằ	Ằằ	Ệệ	Íí	Ỡỡ	Ựự	Ýý	Ỡỡ	Ỡỡ
Ỉỉ	Ỉỉ	Ỡỡ	Ỡỡ	Ỡỡ	Ỡỡ	Ỡỡ	Ỡỡ	Ỡỡ	Ỡỡ

Рисунок 1.10 - всі можливі комбінації діакритичних знаків в'єтнамської мови

Окрім цього, у в'єтнамському письмі використовується обширна система діакритичних знаків, що позначають тональні відмінності. Оскільки діакритики відіграють критичну роль у визначенні тонів, кожна літера, що позначає голосний звук, може доповнюватися одним або двома діакритичними знаками. З допомогою таких знаків загалом можна утворити 134 унікальних літер.

Кожен склад у в'єтнамській мові може бути вимовлений у різних тонах, що відрізняються між собою тривалістю, мелодією, висотою та фонацією. Наприклад, у північно-в'єтнамських діалектах існує 6 тонів (рисунки 1.11, таблиця 1.2) [1]:

Назва	Опис	Діакритичний знак
ngang	середній	-
huyền	низький спадний	`
sắc	високий зростаючий	´
hỏi	середній спадаючий, потім зростаючий	ˆ
ngã	скрипучий високий зростаючий	~
nặng	скрипучий низький спадаючий	˘

Таблиця 1.2 – тони північно-в'єтнамських діалектів

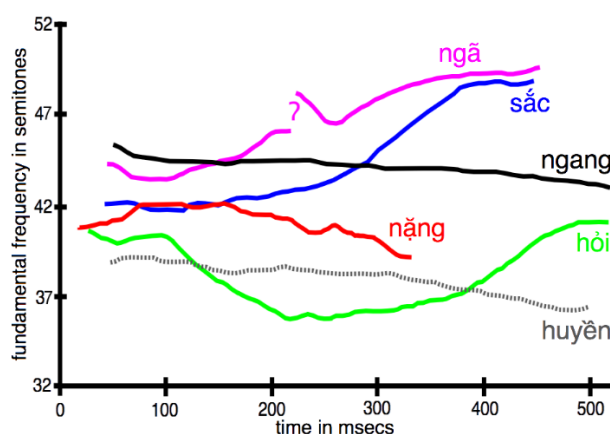


Рисунок 1.11 - візуалізація тонів північно-в'єтнамських діалектів

1.4. Висновки за розділом 1

У даному розділі було розглянуто ключові лінгвістичні особливості 4-х досліджуваних мов та їх системи письма.

Отримані спостереження буде використано у наступних розділах для визначення оптимального способу представлення символів досліджуваних мов у цифровому вигляді та проектування власного редактору методів введення.

РОЗДІЛ 2: Зберігання цифрової інформації різними мовами

2.1. Кодування символів у цифрових системах

Для представлення символів у електронно-обчислювальних машинах використовуються системи кодування символів. В основі роботи більшості таких систем лежить принцип надання кожному символу з репертуару певного числового значення (кодової позиції). Цей підхід дозволяє однозначно ідентифікувати символи, що робить можливим їх зберігання та передачу у цифровому вигляді.

Однією із перших систем цифрового кодування символів, що увійшла в широке використання в комп'ютерах, став стандарт **ASCII** (American Standard Code for Information Interchange), розроблений у 60-х роках XX століття. Система ASCII ґрунтується на англійському алфавіті та надає кодування 128 символам, серед яких є арабські цифри, великі та малі літери англійського алфавіту від а до z та розділові знаки (*рисунок 2.1*).

USASCII code chart

b ₇ b ₆ b ₅ b ₄				b ₃	b ₂	b ₁	b ₀	Column								
b ₇ b ₆ b ₅ b ₄				b ₃	b ₂	b ₁	b ₀	Row								
0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	1	1	NUL	DLE	SP	0	@	P	'	p
0	0	0	0	0	0	1	0	1	SOH	DC1	!	1	A	Q	a	q
0	0	0	0	0	1	0	0	2	STX	DC2	"	2	B	R	b	r
0	0	0	0	0	1	1	1	3	ETX	DC3	#	3	C	S	c	s
0	0	0	0	0	1	0	1	4	EOT	DC4	\$	4	D	T	d	t
0	0	0	0	0	1	1	0	5	ENQ	NAK	%	5	E	U	e	u
0	0	0	0	0	1	1	1	6	ACK	SYN	&	6	F	V	f	v
0	0	0	0	1	1	1	0	7	BEL	ETB	'	7	G	W	g	w
0	0	0	0	1	0	0	0	8	BS	CAN	(	8	H	X	h	x
0	0	0	0	1	0	0	1	9	HT	EM	)	9	I	Y	i	y
0	0	0	0	1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
0	0	0	0	1	0	1	1	11	VT	ESC	+	;	K	[	k	(
0	0	0	1	0	0	0	0	12	FF	FS	,	<	L	\	l	/
0	0	0	1	0	0	0	1	13	CR	GS	-	=	M	]	m	)
0	0	0	1	0	0	1	0	14	SO	RS	.	>	N	^	n	~
0	0	0	1	0	0	1	1	15	SI	US	/	?	O	_	o	DEL

Рисунок 2.1 – таблиця символів ASCII

Довгий час стандарт ASCII був найбільш вживаною системою кодування у Всесвітній мережі. Однак, у нього є один очевидний недолік: ASCII розрахований лише на використання з англійською мовою, а тому не

міг довго залишатися лідером на ринку, який активно переживав період швидкої глобалізації.

Тому, у 2007-у році новим світовим лідером серед систем кодування став стандарт **Юнікод** (англ. **Unicode**), а саме його формат **UTF-8**. Принципи кодування Unicode передбачають можливість кодування 1 112 064 символів, з яких станом на березень 2020-го року Unicode використовує 143 859 – це число покриває 154 системи письма, набори допоміжних символів та емодзі. Оскільки наразі UTF-8 залежно від мови використовує 97-100% вебсторінок Інтернету, саме це кодування надалі розглядатиметься у даній курсовій роботі [6].

Кодування UTF-8 здатне закодувати будь-який символ з допомогою від одного до чотирьох байтів. При цьому, кодування розраховане на максимальну оптимізацію пам'яті. Наприклад, для символів з діапазону U+0000 - U+007F використовується лише 1 байт, а не усі 4. (таблиця 2.1).

Кодові позиції	Байт 1	Байт 2	Байт 3	Байт 4
U+0000 - U+007F	0xxxxxxx			
U+0080 - U+07FF	110xxxxx	10xxxxxx		
U+0800 - U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
U+10000 - U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Таблиця 2.1 – розподіл байтів у стандарті Unicode

Перші 128 символів Юнікоду для числового представлення потребують 1 байт та в точності повторюють всі символи ASCII, що робить Юнікод зворотно сумісним із ASCII. Для кодування наступних 1 920 символів використовується 2 байти – цього достатньо зокрема для покриття майже всієї решти латинських, коптського, грецького, вірменського, та арабського алфавітів, а також іврити, кирилиці, та десятків комбінувальних

діакритичних знаків. 3 байти використовується для покриття решти ВМР (Basic Multilingual Plane – перша з 17-ти площин кодів Юнікоду, що містить практично всі можливі символи у широкому використанні) – сюди входять десятки тисяч символів китайських, японської та корейської мов. Нарешті, 4 байти використовуються для кодування символів з інших площин Юнікоду: менш розповсюджених символів китайських, корейської та японської мов, деяких історичних систем письма, математичних символів та емодзі [6].

2.2. Композиція складних символів у системах кодувань

Окрім уже наявних у каталозі Юнікоду символів, цей стандарт кодування також передбачає створення додаткових через процес **комбінування**.

У Юнікодi комбінування відбувається з допомогою спеціальних **комбінувальних символів**, серед яких найбільш використовуваними зазвичай є діакритичні знаки, хоч цим і не обмежується їх вибір. До будь-якого базового символу Юнікоду можна додати один або декілька комбінувальних – на виході утворюється новий складений символ (*рисунок 2.2*) [6].

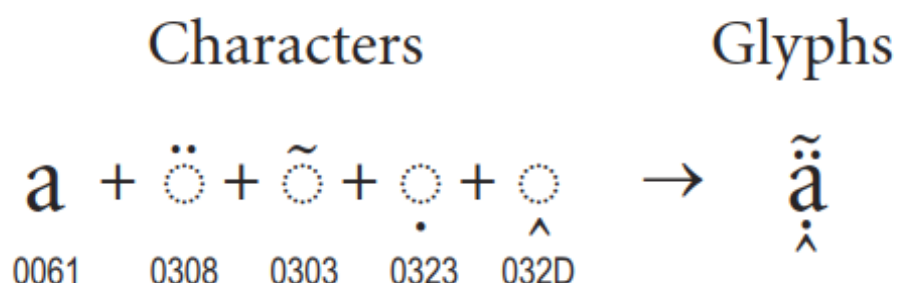


Рисунок 2.2 – приклад комбінування символів в Юнікодi

Важливо, що деякі складені символи, що можна утворити через комбінування, уже наявні у системі кодування як один базовий символ. Наприклад, символ é (U+00E9) може бути представлений через стрічку-комбінацію базової е (U+0065) та гострого наголосу (U+0301).

У Юнікодi є 5 блоків, що повністю відведені для комбінувальних символів (*повний перелік символів наведено у додатку А*):

1. Combining Diacritical Marks (0300–036F)
2. Combining Diacritical Marks Extended (1AB0–1AFF)
3. Combining Diacritical Marks Supplement (1DC0–1DFF)
4. Combining Diacritical Marks for Symbols (20D0–20FF)
5. Combining Half Marks (FE20–FE2F)

Проте, цими блоками комбінувальні символи у Юнікодi не обмежуються – деякі з них знаходяться в одних блоках зі звичайними базовими символами. Наприклад, комбінувальні японські символи дакутен (U+3099) та хандакутен (U+309A) закодовані у блоці японської системи письма хірагана, а блок деванагарі містить комбінувальні голосні знаки, що використовуються у цій індійській системі письма.

2.3. Висновки за розділом 2

В межах цього розділу було розглянути основні принципи роботи системи кодування Юнікод та обґрунтовано оптимальність використання цієї системи для цифрового представлення символів досліджуваних мов.

Здобуті у цьому та попередньому розділах знання буде використано у наступному для об'єктивної оцінки сучасних редакторів методів введення.

РОЗДІЛ 3: Введення цифрових даних для різних систем письма

3.1. Опис поняття та принципів роботи ІМЕ

Редактор методів введення (англ. **IME, Input Method Editor**) – це компонент операційної системи або окрема програма, що дозволяє користувачам вводити символи, які за замовчуванням недоступні у їх пристроях для введення.

Особливо широкого використання ІМЕ набули для введення символів китайських, японської та корейської мов, кожна з яких містить тисячі різноманітних символів, на клавіатурах з латинською розкладкою.

Хоч всі ІМЕ і унікальні у своїй реалізації, часто у них присутні такі два модулі (*проілюстровані на рисунку 3.1*):

- **Компонувальник** – незалежний від контексту парсер, що будує відмінні від ASCII символи через натискання певних кнопок (або їх комбінацій) на клавіатурі. Використовується, наприклад, для написання японських символів хірагана.
- **Конвертер** – контекстно-залежний парсер, що з допомогою словника конвертує фонетичні символи у ідеографічні, наприклад, японські канджі.

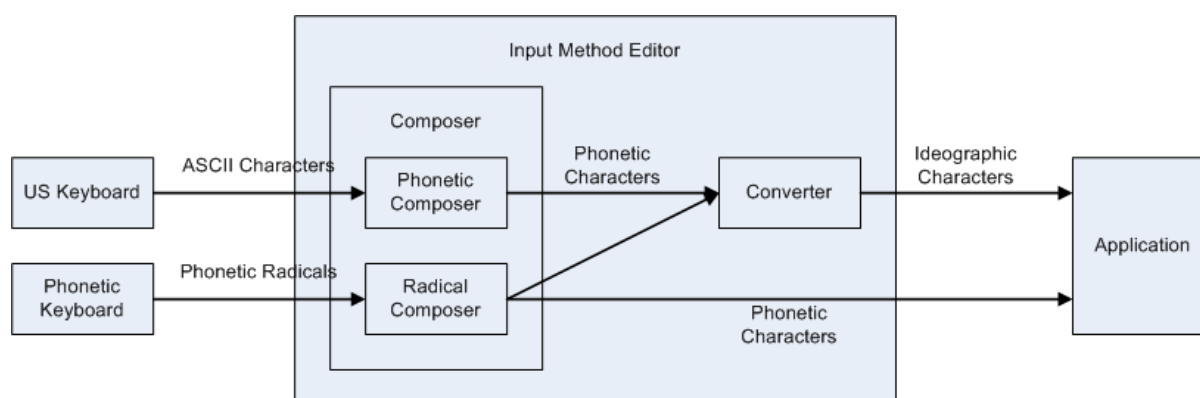


Рисунок 3.1 – можлива структура редактору методів введення

Компонувальники можна розділити на два основні типи:

- **Фонетичні компонувальники** – компонують символи через попереднє введення їх фонетичного представлення.
- **Компонувальники радикалів (ключів)** – будують символи через прив'язку окремих клавіш клавіатури до радикалів – структурних одиниць письма багатьох логографічних мов, з яких складаються більш комплексні символи.

У наступних підрозділах буде розглянуто декілька найбільш поширених типів методів введення, їх варіацій у контекстах різних мов та окремих реалізацій в певних сучасних ІМЕ [7].

3.1.1. Фонетичний метод введення

3.1.2. Огляд та використання методу на прикладі досліджуваних мов

Фонетичні методи введення передбачають ставлення у відповідність символів чи слів однієї мови символам чи словам іншої на основі фонетичної схожості.

Особливої популярності такі методи введення набули у мовах з логографічним письмом, як-от, китайські. Наприклад, для написання китайських символів був створений метод введення **пін'їнь**, побудований на одноіменній латинізації мандаринського діалекту. Метод полягає у введенні латинськими літерами вимови потрібних символів, після чого користувач отримує перелік символів-кандидатів зі схожою вимовою – такий підхід є необхідним для китайських мов, оскільки у них є досить багато слів-омофонів. З наданих символів користувач обирає потрібний, після чого попередньо введене фонетичне представлення замінюється на обраний символ [8].

У японській мові, хоч на відміну від китайських вона і не є цілком логографічною, використовується схожий за принципом метод введення, що називається **ромаджі** й, знову ж таки, запозичує свою назву у системи романізації японських символів. В методі ромаджі написання символів починається із написання латинськими літерами фонетичного представлення символів системи хірагана. Латинські літери замінюються на відповідні символи хірагани в режимі реального часу. Оскільки у японській мові використовуються три різні системи письма (канджі, катакана та хірагана), на цьому етапі користувач може обрати між такими варіантами: залишити символи хірагани, або конвертувати їх у катакану чи канджі [10].

Однак, використання фонетичних методів введення поширене далеко не лише у мовах країн Східної Азії – такий тип введення є універсальним і тому активно використовується й у інших мовах, як-от арабська чи гінді.

3.1.3. Реалізація методу в різних ІМЕ

Одним із найпопулярніших ІМЕ для веб-браузерів є **Google Input Tools** – редактор методів введення, доступний у вигляді розширення браузеру Google Chrome. Google Input Tools надає якісні способи введення для десятків мов.

Зокрема, пін'їнь: користувач вводить пін'їнь символу та отримує символи-пропозиції. ІМЕ надає зручне керування меню пропозицій – є можливість прогортати усі варіанти та обрати потрібний з допомогою натискання клавіші миші або відповідного порядкового номеру на клавішах клавіатури (рисунк 3.2).

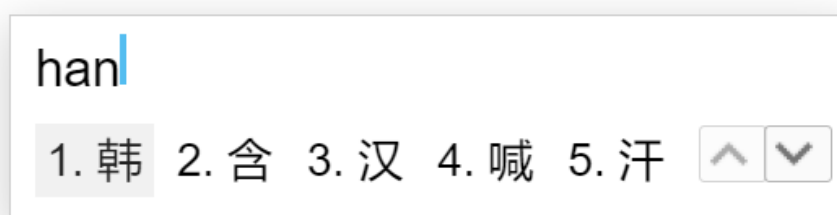


Рисунок 3.2 – піньїнь в Google Input Tools

Даний редактор надає і метод введення японської мови ромаджі. У цьому ІМЕ вимова символів хіраґани, виражена через латинські літери, одразу конвертується у хіраґану та виводиться на екран. Після натискання клавіші «пробіл» на клавіатурі, на екрані з'являються варіанти символів та їх сполучень з подібною вимовою – варіанти наявні у всіх трьох системах письма, тому користувач має змогу вибрати потрібну (рисунки 3.3).

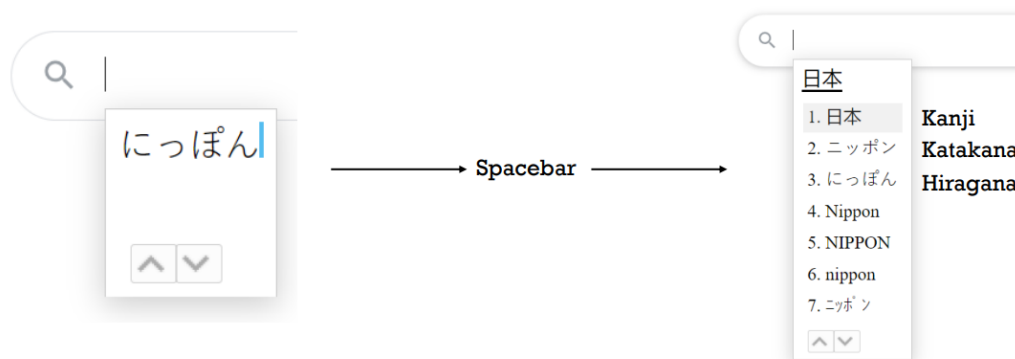


Рисунок 3.3 – ромаджі в Google Input Tools

У додатку Б також наведені приклади реалізації фонетичних методів введення у інших ІМЕ.

3.2. Методи введення, що опираються на розкладку клавіатури

3.2.1. Огляд та використання методів на прикладі досліджуваних мов

У поточному підрозділі розглядається декілька методів введення, всі з яких, хоч і є різними за принципом дії, але залежать від безпосередньої прив'язки клавіш клавіатури до символів необхідної мови.

Яскравим прикладом такого типу методів введення є **вубі** – метод введення китайських мов, що для введення символів використовує їх структуру, а не фонетичне представлення, як, наприклад, пін'їнь. Метод вубі передбачає розподіл клавіатури на 5 регіонів (рисунки 3.4), кожен з яких відповідає за певний тип рисок (структурних складових китайських символів) [9]:

1. Горизонтальні (一)
2. Вертикальні (丨)
3. Спадні справа наліво (丿)
4. Точкові або спадні зліва направо (丶)
5. Гачкуваті



Рисунок 3.4 – розподіл символів у методі вубі

Метод вубі, хоч часто і складний для нових користувачів, вважається вкрай ефективним: практично будь-який символ з допомогою цього методу

може бути написаний щонайбільше з допомогою 4-ох натискань на клавіші клавіатури.

Дещо простішим є японський аналог методу введення такого типу: окрім фонетичного ромаджі, для написання японських символів часто використовується звичайна клавіатура, в якій окрім латинських символів кожна клавіша також відповідає за певний символ хірагана (*рисунок 3.5*). Оскільки система хірагана загалом нараховує 46 базових символів, відобразити їх з допомогою стандартної клавіатури не є проблемою [10].



Рисунок 3.5 – клавіатура з розкладкою хірагана

Аналогічні методи введення, що опираються на власну розкладку клавіатури, існують для більшості загальноновживаних мов, в яких порівняно невелика кількість символів у письмі дозволяє задовольнити потребу введення таким чином.

Наприклад, для введення арабської мови використовується спеціальна розкладка (та декілька її варіацій) клавіатури з арабськими символами (*рисунок 3.6*).

~	!	@	#	\$	%	^	&	*	(	)	-	+	← Backspace	
Tab	ض	ص	ث	ق	ف	غ	ع	ه	خ	ح	ج	د		
Caps Lock	ش	س	ي	ب	ل	أ	ت	ن	م	ك	ط	Enter		
Shift	ئ	ء	ؤ	ر	لا	آ	ة	و	ز	ظ	؟	Shift		
Ctrl	Win Key	Alt									Alt Gr	Win Key	Menu	Ctrl

Рисунок 3.6 – арабська розкладка клавіатури

3.2.2. Реалізація методу в різних IME

Google Input Tools дозволяє використовувати метод вубі для введення китайських символів й візуально реалізація досить схожа на пін'їнь. Насправді ж, алгоритм введення дещо інший: користувач вводить риси з допомогою 5-и різних регіонів клавіатури і таким чином буде необхідний складений символ. У результаті, на екран виводиться перелік символів-кандидатів, адже схожі комбінації рисок можуть надавати декілька результатів (рисунок 3.7).



Рисунок 3.7 – реалізація методу вубі в Google Input Tools

В той же час, мобільна ОС IOS, наприклад, надає зручний редактор для введення символів хірагани (рисунок 3.8).



Рисунок 3.8 – редактор для введення хіраґани в IOS

У додатку Б наведені приклади реалізації аналогічних методів введення у деяких інших ІМЕ.

3.3. Інші методи введення

3.3.1. Огляд та використання методів на прикладі досліджуваних мов

Окрім найбільш поширених методів введення, що були розглянуті у двох попередніх підрозділах, існують десятки інших, більш нішевих або просто менш ефективних, а тому і не таких популярних.

Серед таких методів можна виділити, наприклад, рукописний – метод введення, що передбачає написання слів та символів у спеціальному вікні з допомогою курсора на комп'ютері або пальців руки на смартфонах [11]. Попри те, що такий метод введення з першого погляду може здатися найприроднішим (адже він повторює написання тексту від руки), часто він не є найбільш ефективним і програє більш практичним методам введення.

Тим не менше, сьогодні його можна зустріти у більшості ІМЕ, доступним для будь-яких підтримуваних мов.

З унікальних для певної мови методів можна також виділити метод введення в'єтнамської мови **Телекс**. Оскільки в'єтнамська мова великою мірою покладається на комплексну систему діакритичних знаків, основною задачею телексу є максимально ефективно додавати ці знаки до базових символів. Виразити тони в'єтнамської мови через діакритичні знаки телекс дозволяє з допомогою певних допоміжних символів: користувач спершу вводить слово з базовим голосним символом, а опісля вводить додатковий символ, що вказує, який діакритик має бути використаний (*таблиця 3.1*).

Тон	Клавіші	Приклад введення	Результат
Ngang (level)	z / нічого	ngang	ngang
Huyền (falling)	f	huyeenf	huyền
Sắc (rising)	s	sawcs	sắc
Hỏi (dipping-rising)	r	hoir	hỏi
Ngã (rising glottalized)	x	ngax	ngã
Nặng (falling glottalized)	j	nawngj	nặng

Таблиця 3.1 – введення тональних діакритиків з допомогою телексу

Для нетональних діакритиків використовується аналогічна система (*таблиця 3.2*).

Символ	Клавіші	Приклад введення	Результат
ă	aw	trang-w	trăng
â	aa	can-a	cân
đ	dd	d-dau-a	đâu
ê	ee	d-dem-e	đêm
ô	oo	nho-o	nhô
ơ	ow / [	mo-w / m[	mơ
ư	uw / w /]	tu-w / tw / t]	tư

Таблиця 3.2 – введення нетональних діакритиків з допомогою телексу

3.3.2. Реалізація методів у різних IME

Google Input Tools надає рукописний метод введення практично для всіх наявних у редакторі мов. Для цього користувачу надається спеціальне віконце, у якому символи чи слова малюються від руки. Аналогічно з попередніми методами, доступний також перелік символів чи слів-кандидатів та зручна навігація по них (рисунки 3.9).



Рисунок 3.9 – рукописне введення японської, арабської та української мов в Google Input Tools

В цьому ж ІМЕ присутня також реалізація методу телекс. Даний метод не потребує додаткового графічного інтерфейсу, оскільки весь функціонал побудований на комбінаціях клавіш, але для додаткової зручності ІМЕ надає віртуальну клавіатуру (рисунок 3.10).

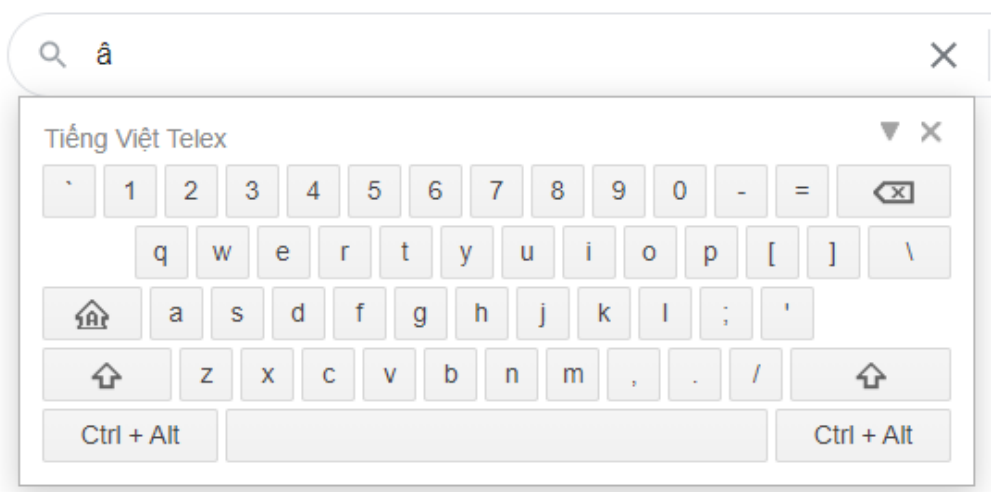


Рисунок 3.10 – телекс в Google Input Tools

Додаткові скріншоти реалізації методів з цього підрозділу у інших редакторах можна проглянути у додатку Б.

3.4. Висновки за розділом 3

У третьому розділі було розглянуто загальні принципи роботи ІМЕ та досліджено сучасні методи введення, наявні для досліджуваних мов.

Проведено огляд різних реалізацій цих методів у доступних на ринку ІМЕ.

З огляду на кількість та ефективність розглянутих ІМЕ, можна зробити висновок, що наразі потреба ринку у зручних методах введення задоволена.

РОЗДІЛ 4: Реалізація та огляд власного ІМЕ

4.1. Опис та обґрунтування формату запропонованого рішення

У попередньому розділі було розглянуто сучасні редактори методів введення та проаналізовано їхню придатність для використання з досліджуваними мовами. Було зроблено висновок, що для введення символів окремих мов наразі існує достатньо реалізованих ІМЕ й ринок не потребує аналогів.

Тому, в межах даної курсової роботи було прийнято рішення не створювати редактор, прив'язаний до конкретної мови чи декількох мов. Натомість, застосунок покликаний продемонструвати можливість розширення стандартної розкладки клавіатури власними символами.

Розроблюваний веб-редактор орієнтований на ширше використання з будь-яким набором символів, переданим користувачами, що інтегруватимуть його у власні системи. Каталог символів передається у застосунок у вигляді нового шрифту (в файлі формату .ttf або .otf), в якому кожен модифікований символ обов'язково прив'язаний до певного кодування з таблиці Юнікоду. Для коректного відображення символів шрифт також має бути довантажений на веб-сторінку. Опісля, новостворені символи вводяться з допомогою спливаючого вікна, в якому дається вибір з декількох варіацій базового символу, введенного кінцевим користувачем.

4.2. Технології та інструменти, обрані для реалізації

У реалізації застосунку було використано такі технології:

- **JavaScript** – для основного функціоналу застосунку: роботи з символами та реалізації інтерактивного вікна редактора.

- **HTML/CSS** – для побудови демо-сторінки, редактору методів введення та надання усім компонентам потрібних стилів.
- **Glyphs** – macOS-додаток для побудови шрифтів, дозволяє створювати власні символи та додавати їх у каталоги.
- **NPM** – система-менеджер пакунків для мови програмування JavaScript та її платформи **Node.js**. Була використана для інтеграції сторонньої бібліотеки з метою оптимізації роботи застосунку.

4.3. Архітектура та реалізація застосунку

До початку розробки самого застосунку, було підготовано оригінальний шрифт на основі сучасного шрифту Arial. Основні латинські символи було залишено без змін – натомість, було змінено зовнішній вигляд 208-и складених символів, наприклад, таких, що складаються з базових латинських символів та діакритичних знаків (*рисунк 4.1*).

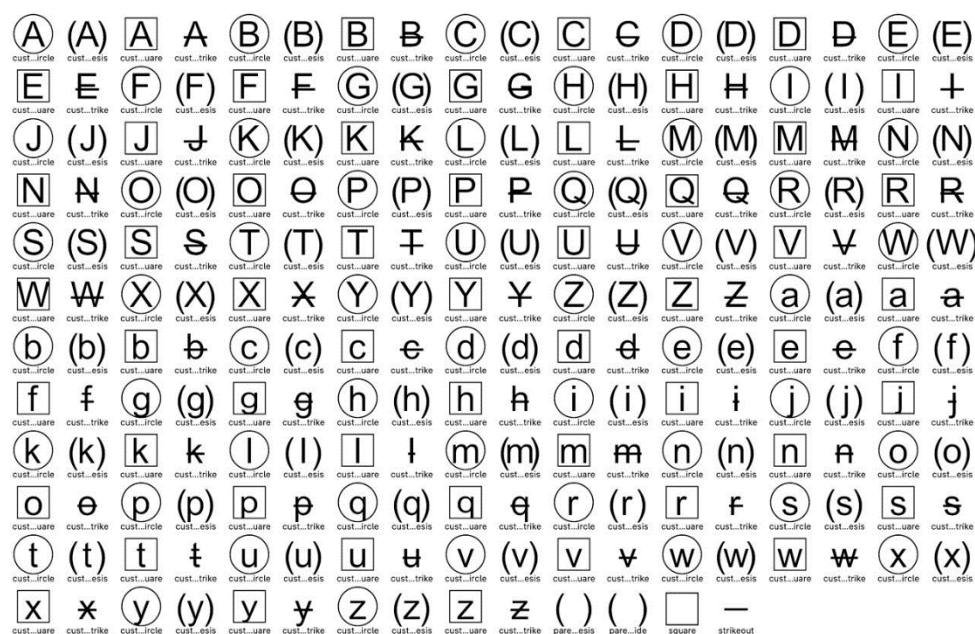


Рисунок 4.1 – повний перелік новостворених символів

Новий шрифт було збережено в .ttf-файл, який після цього було завантажено на веб-сторінку з допомогою відповідних конфігурацій CSS.

У JavaScript-кодi, що відповідає за основний функціонал застосунку, було створено об'єкт-словник, в якому у відповідність усім базовим літерам латинського алфавіту поставлено у відповідність по 4 варіації символу, що виражені через певні коди системи Юнікод (рисуюнок 4.2).

```
const charDict = {
  "A": ["\u00C0", "\u00C1", "\u00C2", "\u00C3"],
  "B": ["\u01E0", "\u01E4", "\u01E6", "\u01D2E"],
  "C": ["\u0106", "\u0108", "\u010A", "\u010C"],
  "D": ["\u010E", "\u01C4", "\u01C5", "\u01F1"],
  "E": ["\u00C8", "\u00C9", "\u00CA", "\u00CB"],
  "F": ["\u01E8", "\u0210", "\u0213B", "\u024BB"],
  "G": ["\u011C", "\u011E", "\u0120", "\u0122"],
  "H": ["\u0124", "\u021E", "\u01D34", "\u01E22"],
  "I": ["\u00CC", "\u00CD", "\u00CE", "\u00CF"],
  "J": ["\u0132", "\u0134", "\u01C7", "\u01CA"],
  "K": ["\u0136", "\u01E8", "\u01D37", "\u01E30"],
  "L": ["\u0139", "\u013B", "\u013D", "\u013F"],
  "M": ["\u01D3", "\u01E3E", "\u01E40", "\u01E42"],
  "N": ["\u00D1", "\u0143", "\u0145", "\u0147"],
  "O": ["\u00D2", "\u00D3", "\u00D4", "\u00D5"],
}
```

Рисуюнок 4.2 – фрагмент об'єкту, що зберігає варіації символів

При завантаженні веб-сторінки з застосунком використовується метод **addEventListener()** інтерфейсу **EventTarget** для відстеження фокусування на текстових полях сторінки (рисуюнок 4.3).

```
function initIME() {
  document.addEventListener('focusout', (e) => {
    setTimeout(() => {
      if (document.activeElement.id !== "ime-container" && document.activeElement !== e.target) {
        disableIME();
      }
    }, 1);
  });
  document.addEventListener('focusin', (e) => {
    if (e.target.tagName === "TEXTAREA") {
      focusedInputElement = e.target;
      setTimeout(() => {
        enableIME();
      }, 1);
    }
  });
}
```

Рисуюнок 4.3 – відстеження фокусу на текстових полях

У випадку фокусування на текстовому полі викликається метод **enableIME()** (рисунк 4.4), у якому:

- ініціалізуються DOM-елементи, відповідальні за відображення редактора.
- створені елементи вводяться в DOM-дерево й відбувається візуальна оптимізація їх розташування.
- на сфокусоване текстове поле накладається відстеження декількох можливих типів подій (інтерфейс **Event**), що спричиняються різними типами введення.

```
function enableIME() {
    disableIME();
    focusedInputElement.addEventListener("input", inputHandler);
    focusedInputElement.addEventListener('keydown', inputHandler);
    focusedInputElement.addEventListener('mousedown', inputHandler);
    focusedInputElement.addEventListener('select', inputHandler);

    let imeContainer = document.createElement("div");
    imeContainer.id = "ime-container";
    imeContainer.setAttribute("tabindex", "0");

    let imePlaceholderContainer = document.createElement("div");
    imePlaceholderContainer.id = "ime-placeholder-container";
    imePlaceholderContainer.innerText = "Keep writing!";
    imeContainer.appendChild(imePlaceholderContainer);

    focusedInputElement.parentNode.insertBefore(imeContainer, focusedInputElement.nextSibling);
    positionIME();
}
```

Рисунок 4.4 – метод enableIME(), що відповідає за початкову конфігурацію редактора

Нарешті, метод **inputHandler()** (рисунк 4.5) відповідає за обробку подій, викликаних фактом введення символів у сфокусоване текстове поле. Метод по-різному обробляє різні типи подій, але основний алгоритм у всіх випадках складається з таких кроків:

1. Зчитування базового символу, введеного користувачем.
2. Пошук базового символу у словнику **charDict** та отриманням з нього усіх наявних варіацій.
3. Заповнення вікна редактора отриманими складеними символами.

```

if (charDict.hasOwnProperty(lastInputChar)) {
  for (let baseChar of charDict[lastInputChar]) {
    let charElement = document.createElement("div");
    charElement.className = "list-item";
    charElement.onclick = () => {
      imeClickHandler(charElement, oldSelectionIndex);
    }
    let charElementInner = document.createElement('div');
    charElementInner.className = "list-item-inner";
    charElementInner.textContent = baseChar;
    charElement.appendChild(charElementInner);
    document.getElementById("ime-container").appendChild(charElement);
  }
}

```

Рисунок 4.5 – фрагмент методу inputHandler(), що відображає ключовий алгоритм обробки введення

Зрештою, якщо базовий символ було знайдено у словнику, вікно редактора заповнюється складеними символами, що базуються на базовому. Ці символи доступні у вікні у вигляді інтерактивних кнопок, при натисканні на які базовий символ у текстовому полі замінюється на відповідний складений. Як альтернатива натисканню на кнопки, редактор також дозволяє зчитувати комбінації клавіш Alt + (1-4) для форматування тексту з допомогою клавіатури, а не миші.

4.4. Демонстрація роботи застосунку

При завантаженні демо-сторінки для редактора перед користувачем з'являється текстове поле з наперед заповненим текстом (рисунк 4.6).

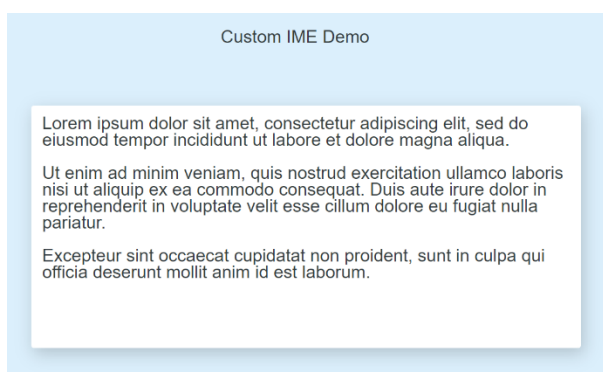


Рисунок 4.6 – початковий вигляд демо-сторінки

При фокусуванні на текстовому полі, під текстовою кареткою з'являється вікно редактора – спочатку воно містить контекстно-незалежний текст-заклик (*рисунок 4.7*).

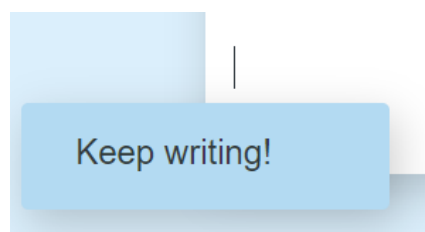


Рисунок 4.7 – вікно редактора за відсутності введенного символу в словнику

Щойно користувач вводить символ, який є у переліку базових символів у словнику, у вікні редактора у вигляді кнопок з'являються всі можливі варіації базового символу. При натисканні на будь-який з варіантів надрукований користувачем базовий символ замінюється відповідним складеним символом. Окрім натискання клавішою миші, обрати необхідний складений символ можна також з допомогою комбінації клавіші Alt + та цифрових клавіш від 1 до 4, залежно від варіанту. (*рисунок 4.8*).

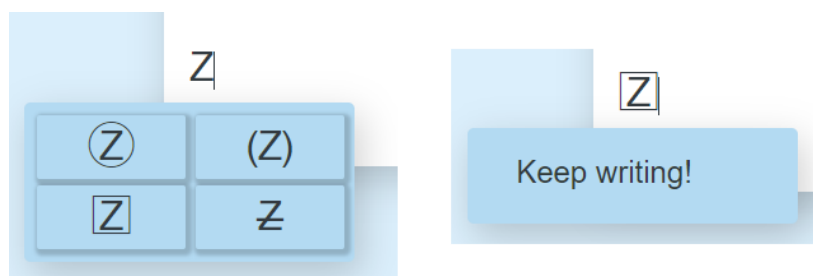


Рисунок 4.8 – вікно редактора при введенні базового символу

Окрім цього, редактор надає можливість формувати уже надрукований текст з допомогою тих же допоміжних форм. Для цього, користувачу потрібно виділити певну частину тексту, після чого у вікні

редактора з'являться аналогічні позначення форм, які можна застосувати до виділення (рисунки 4.9).

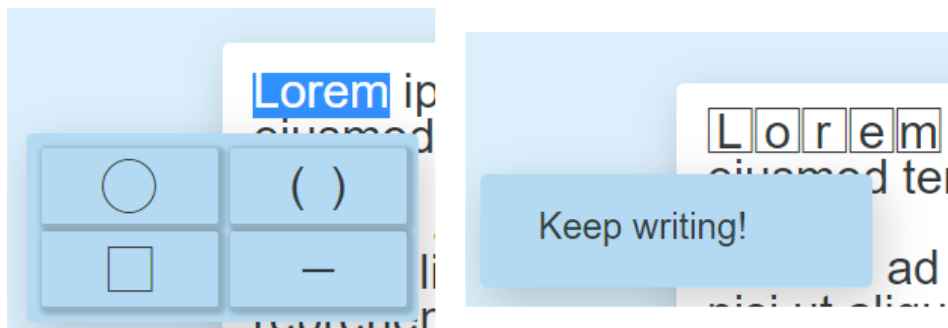


Рисунок 4.9 – вікно редактора при виділенні уже надрукованого тексту

4.5. Висновки за розділом 4

З допомогою знань, отриманих у попередніх розділах даної курсової роботи, було розроблено редактор методів введення, що демонструє можливий підхід введення в використання власних символів.

Проведено огляд внутрішніх алгоритмів застосунку та продемонстровано роботу редактору на практиці.

Висновки

Поширене використання електронно-обчислювальних машин та всесвітньої мережі Інтернет у сучасному світі створює необхідність зручних інструментів для взаємодії з цими технологіями, серед яких є і редактори методів введення.

У межах даної роботи було досліджено декілька ключових аспектів, що є важливими у розробці ІМЕ, а саме:

- лінгвістика мов, для яких створюється редактор;
- способи цифрового представлення інформації, записаної цими мовами;
- наявні підходи до введення символів цих мов з допомогою поширених пристроїв;

Результати дослідження було використано для створення власного редактору методів введення, що бере до уваги потреби ринку та надає актуальне вирішення для ситуацій, де необхідна більша гнучкість у наборах символів, з якими працює кінцевий користувач.

Завдяки своїй структурі, застосунок може бути легко адаптований під потреби різних користувачів. Водночас, у створеного ІМЕ є перспективи подальшого розвитку: планується додати ширший функціонал для форматування тексту, провести роботу з графічного покращення доступності застосунку та у вигляді розширення браузеру зробити його універсальним для будь-яких веб-сторінок.

Література

1. Ethnologue (22nd ed., 2019). [Електронний ресурс]
<https://www.ethnologue.com/language/cmn>
2. Nationalencyklopedin, "Världens 100 största språk 2010" (The World's 100 Largest Languages in 2010)
3. "Arabic – Ethnologue". Ethnologue. Simons, Gary F. and Charles D. Fennig (eds.). 2018. Ethnologue: Languages of the World, 21st edition.
4. "East Asian Languages and Chinese Characters". Pinyin.info. [Електронний ресурс] http://www.pinyin.info/readings/texts/east_asian_languages.html
5. Kazuko Nakajima (2002). Learning Japanese in the Network Society. University of Calgary Press. p. xii. ISBN 978-1-55238-070-3.
6. "The Unicode Standard, Version 13.0– Core Specification Appendix C" (PDF). Unicode Consortium. Retrieved 2020-03-11. [Електронний ресурс] <https://www.unicode.org/versions/Unicode13.0.0/appC.pdf>
7. Input Method Editor API, W3C Working Group, Note 24 May 2016.
[Електронний ресурс] <https://www.w3.org/TR/ime-api/>
8. "About Pinyin input (and other methods): Chinese typing basics and background :: Pinyin Joe". [Електронний ресурс] www.pinyinjoe.com
9. Wicentowski, Joe (1996), Wubizixing for Speakers of English.
10. Google Japan. "Godan キーボードとはなんですか?", 2021
[Електронний ресурс]
<https://support.google.com/ime/japanese/answer/2700298?hl=ja>
11. A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber. A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, 2009.

Додаток А (довідниковий)

Блоки Юнікоду, повністю відведені під комбінувальні символи

Combining Diacritical Marks																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+030x	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+031x	◌̠	◌̡	◌̢	◌̣	◌̤	◌̥	◌̦	◌̧	◌̨	◌̩	◌̪	◌̫	◌̬	◌̭	◌̮	◌̯
U+032x	◌̰	◌̱	◌̲	◌̳	◌̴	◌̵	◌̶	◌̷	◌̸	◌̹	◌̺	◌̻	◌̼	◌̽	◌̾	◌̿
U+033x	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿	◌̿
U+034x	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	CGJ
U+035x	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+036x	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟

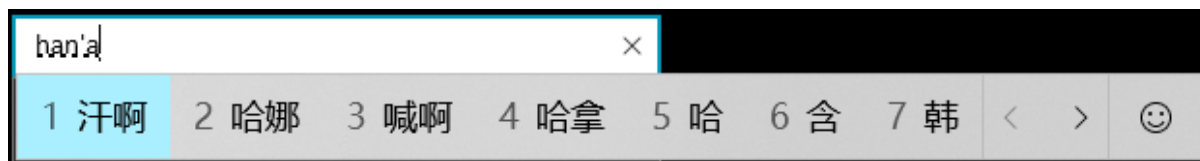
Combining Diacritical Marks Extended																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+1ABx	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+1ACx	◌̐															
U+1ADx																
U+1AEx																
U+1AFx																

Combining Diacritical Marks Supplement																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+1DCx	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+1DDx	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+1DEx	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟
U+1DFx	◌̐	◌̑	◌̒	◌̓	◌̔	◌̕	◌̖	◌̗	◌̘	◌̙		◌̛	◌̜	◌̝	◌̞	◌̟

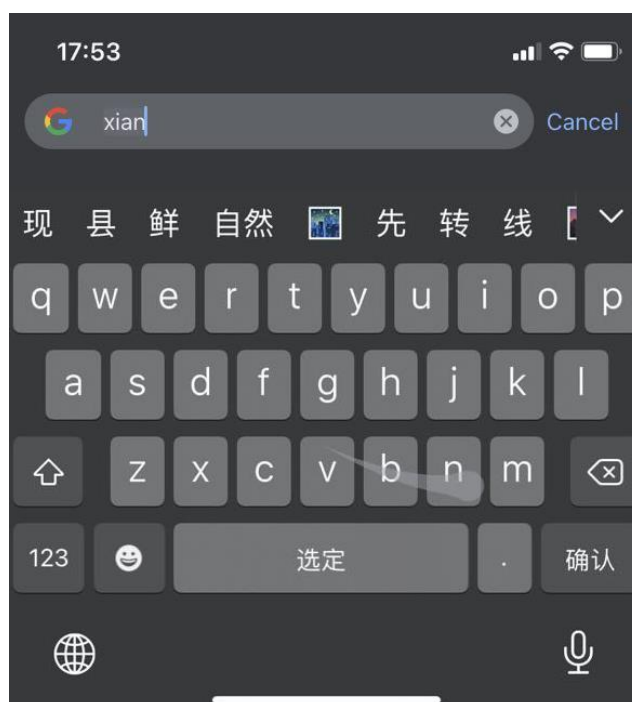
Додаток Б (довідниковий)

Графічні приклади реалізації методів введення в різних ІМЕ

Пін'їнь в ОС Microsoft Windows:



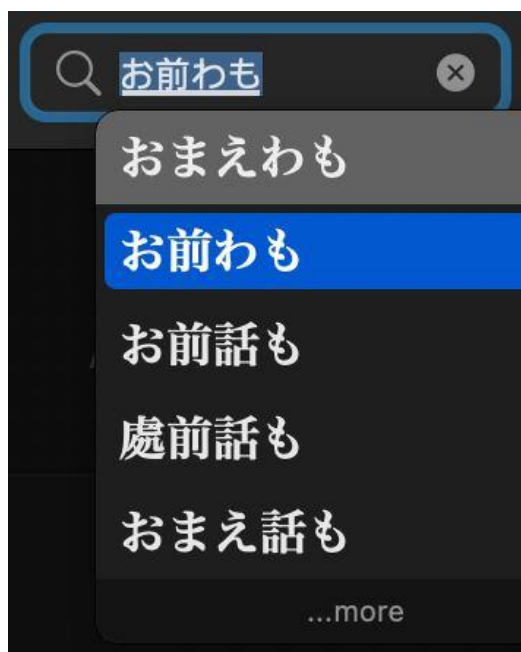
Пін'їнь в ОС IOS:



Пін'їнь в ОС macOS:



Ромаджі в ОС macOS:



Ромаджі в ОС IOS:



Вибі в ОС IOS:



Арабська розкладка клавіатури в ОС IOS:



Телекс в ОС IOS:

