Corpus-Based, High-Productivity, Global Language Input Systems for Manipulation-Challenged Users

Dan Mailman

ABSTRACT

µLex "mu-Lex" is a global language lexical input user interface system.
This study

1. Introduces ***µLex*** via sample application for English word and phrase selection.
2. Illustrates non-letter ***µLex*** applications via sample application for Mandarin.
3. Describes a novel method for adapting ***µLex*** to manipulation-challenged users.
4. Outlines an algorithm generating ***µLex*** applications from language corpus frequency lists.
5. Discusses lexical input productivity ***vis-à-vis*** the sample applications.

KEYWORDS

input, input methods, manipulation challenged, corpus processing, computer, computer input.

1. Introduction

Rheumatoid arthritis and other manipulation-challenging conditions make knuckle-based input activity (keyboarding, mouse-clicking) problematic, especially for users who produce longer, more complex, or more specialized documents. This study introduces aspects of the ***µLex*** global language input system that facilitate the development of applications alleviating such difficulties for **manipulation-challenged** (**MC**) users.

µLex's ***raison d'être*** is user-selection of text items for transmission to computer applications (e.g., word processors). Although ***µLex*** can transmit individual characters, letters, symbols, &c., this study focuses on increasing productivity in the selection of **lexemes** - text versions of semantically-bound, possibly multi-word (e.g., for English) or multiple graphically complex character (e.g., for Mandarin) sequences of unicode characters. Example common internet lexemes: "I love you" and "奥力给" (/àolìgěi/, "awesome(slang)"). Lexemes can be viewed as sequences of (unicode) **lexons**.

Recent increases in the variety, availability, and quality of lexeme frequency lists produced from global language corpora make it feasible to rapidly create and experiment with ***µLex*** applications to increase productivity in the creation and editing of large, complex, or specialized documents. Among the goals in the development of ***µLex*** is increasing productivity measured as **gestural efficiency** (**GE**). The current study quantifies GE using **lexemes per gesture** (**LPG**) and relates it to selection of longer lexemes with fewer (e.g.,) keyboard or mouse gestures[1].

One arena of experimentation is applications that increase productivity for MC users. A subset of MC users is limited by difficulty or pain actuating the keys on keyboards or buttons on mice. This study introduces a family of lexical input solutions and an associated novel input method for this user subset.

Subsequent sections:

1. Describe a sample ***µLex*** application for English lexemes.
2. Describe a sample ***µLex*** application for Mandarin lexemes.
3. Detail a novel method for ***µLex*** usage by manipulation challenged users.
4. Use Mandarin to sketch algorithmic generation of ***µLex*** applications from frequency lists.

[1] Future studies will explore measuring GE in lexons (rather than lexemes) per gesture.

5. Discuss measuring input software efficiency.

Current and future related research is indicated in footnotes.

2. English Lexeme Sample Application

µLex applications send user input to applications on windowing **operating systems** (**OS**s). Such OSs often display several open applications, but only one application at a time has **input focus**; the **focus application** is the one that receives input from keyboards.

To illustrate the ***µLex*** user interface, consider an example application - ***µWords*** - for selecting English lexemes. The goals of ***µWords*** are (1) to increase the number of characters input for the fewest number of gestures, and, with usage, (2) to minimize time and effort spent on the increased gestural productivity.

2.1. Initial Display

µWords presents a **grid** (Figure 1) of **cells** whose visual contents and functionalities change depending on user interaction. There are three cell types (top-to-bottom, left-to-right):

1. Initial **summary cells** (green) summarize available choices and enable OS options, such as 'move' and 'exit'.
2. Internal **lexeme cells** (white, cyan) enable lexeme, letter, character, and symbol selection.
3. Final **options cells** (blue) enable general lexeme options, such as uppercase for English.

µWords	a (a…)	and (b…)	are (c…)
as (d…)	at (e…)	be (f…)	but (g…)
by (h…)	for (i…)	from (j…)	had (k…)
have (l…)	he (m…)	his (n…)	i (o…)
in (p…)	is (q…)	it (r…)	not (s…)
of (t…)	on (u…)	's (v…)	that (w…)
the (x…)	they (y…)	this (z…)	to ('…)(.?!)
was ←(a-z)	with ←(0-9)	you ↲(@#$)	Options

*Figure 1: Default **µWords** for BNC Selects 44.62% of Lexemes and 44.27% of Text*

In addition to maximizing GE, another strategy ***µWords*** uses to increase productivity is presenting most-frequent lexemes in frequency order to reduce visual scan time and cognitive effort.

2.2. Actuation, Selection, Transmission, Display

Viewing the input process as a series of cyclic interactions, each input cycle (ideally) occurs in a low single-digit number of seconds. The entities involved in the input cycle are users, operating systems, ***µWords*** (or other ***µLex*** applications), and focus applications (Figure 2). With reference to the input cycle, this study applies specific meanings to several words in general use; in particular: **presentation**, **transmission**, **selection**, **actuation**, **option**, and **choice**.

Presentation refers to the contents of the grid's cells at any time. In ***µWords***' initial (and default) presentation, each lexeme cell presents a lexeme (**lex**, black) and a single qwerty symbol (**key**, red). The lex comprise corpus-specific most-frequent lexemes. Default grid keys include all qwerty/alphabetic symbols that prefix corpus lexemes.

Transmission is the sending of (generally, lexical) items from ***µWords*** to focus applications. **Selections** are items with potential to be transmitted to focus applications. These are specified by the applications themselves and can be lexemes, characters, symbols, numbers, commands, control characters, &c. ***µWords*** selection transmission is a supplement to - not (necessarily) a replacement for - device/OS transmission.

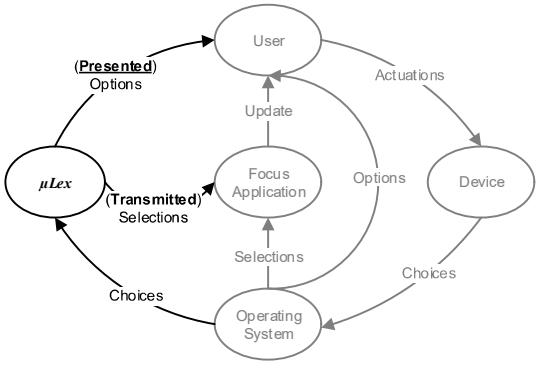

*Figure 2: **µLex** Input Cycle*

Actuation refers to users' device-specific gestures. ***µLex*** actuation is via hardware such as keyboards, mice, touchscreens, as well as specialized environment- or user-specific devices (Figure 3).

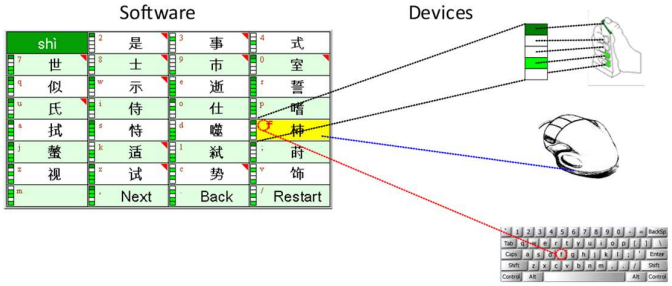

Figure 3: µLex is device-agnostic.

For any device, ***µWords*** distinguishes 4 classes of actuation, each of which may be applied to a single cell at any one time. Although designed to be actuated by a variety of hardware devices, ***µWords*** defaults to actuation via keyboards or mice. To keep discussion on point for this study, mice are taken as representative of hardware input devices[2]. For mice, the four different actuations are:

- Left single click (**LSC**)
- Left double click (**LDC**)
- Right single click (**RSC**)
- Right double click (**RDC**)

µWords cells indicate user **options**. Example lexeme cell options[3] and their associated actuations are:

- LSC selects lex followed by a space to the focus application.
- LDC selects lex without a space and presents punctuation options (Figure 4).
- RSC presents key-prefixed lexemes and multi-character keys (Figure 5).
- RDC transmits sequences of characters indicated by keys.

µPunct	␣	,␣	.␣␣
.␣	.	;␣	:␣
?␣␣	!␣␣	↵	→
'	"	`	␣-␣
←	,	;	:
?	!	-	
		Lock	µWords

Figure 4: Punctuation Choices

C...	c (ca...)	ca (cb...)	called (cc...)
came (cd...)	can (ce...)	car (cf...)	care (cg...)
case (ch...)	cases (ci...)	centre (cj...)	century (ck...)
certain (cl...)	certainly (cm...)	change (cn...)	child (co...)
children (cp...)	church (cq...)	city (cr...)	class (cs...)
clear (ct...)	come (cu...)	committee (cv...)	community (cw...)
company (cx...)	control (cy...)	could (cz...)	council
country	course	court	µWords

Figure 5: C-Prefix Lexeme Choices

[2] Keyboard and specialized device actuation are the subjects of forthcoming studies.
[3] Options presented by summary and options cells are the subjects of forthcoming studies.

A ***µWords*** **choice** is a single cell among those presented by the grid. Since there are 32 choices and four possible actuations for each, there are up to 128 possible functionalities available on the grid at any given time. Actuated choice functionalities for lexeme cells generally involve changes in grid presentation, selection transmission, or both.

Example interactions: on the yellow-highlighted ***µWords*** cell in Figure 1:

- LSC selects (and transmits) /are/+<space>
- LDC selects /are/ and presents the grid in Figure 4
- RSC presents the grid in Figure 5
- RDC selects the single lexon /c/

3. Mandarin Lexeme Sample Application

Analogous to ***µWords*** for English lexemes, ***µ*短语** (‘mu’+/duǎnyǔ/, “phrase”) is designed to increase GE selecting Mandarin lexemes. Like ***µWords***, ***µ*短语** initially displays most frequent lexemes in frequency order (Figure 6). Since meta-lexeme considerations (like capitalization) are different for English and Mandarin, actuations associated with choices are different. The principal meta-lexeme issue addressed by ***µ*短语** is the phenomenon of Mandarin 汉字/漢字 (/hànzì/, “Chinese characters”) having both traditional and simplified form. So, actuations for ***µWords*** and ***µ*短语** have different functionalities:

- LSC/TSP transmits 简化字 (/jiǎnhuàzì/, “simplified Chinese characters”) of the presented lex
- LDC/TLP transmits 繁體字 (/fántǐzì/, “traditional Chinese characters”) of the presented lex
- RSC/PSP presents a grid of key-prefixed lexemes (Figure 7)
- RDC/PLP transmits the (NB: non-lexon) key alphanumeric

µ短语	A 的	B 我	C 了	D 你	E 是	F 在	G 有
H 不	I 个	J 好	K 就	L 啊	M 人	N 都	O 也
P 要	Q 这	R 去	S 说	T 和	U 很	V 会	W 看
X 吧	Y 自己	Z 到	1 来	2 给	3 上	4 想	选项

Figure 6: Default µ短语 for LCMC Selects 27.28% of Lexemes and 23.39% of Text

s-	A 什么	B 时候	C 手	D 死	E 谁	F 时	G 事
H 生活	I 时间	J 睡	K 水	L 世界	M 所以	N 送	O 岁
P 双	Q 睡觉	R 事情	S 少	T 啥	U 是不是	V 会	W 生
X 身	Y 所有	Z 虽然	1 所	2 伤	3 神	4 傻	选项

Figure 7: Presentation for Selecting Pīnyīn Key-Prefix Lexemes

The general strategy for ***µ*短语** is to select lexemes’ qwerty/pinyin prefixes with (blue) lowercase letters and transmit displayed lexemes with uppercase letters.

For example, the initial displayed grid (Figure 6) indicates that uppercase <S> transmits “说” (/shuō/,”speak”) and lowercase <s> updates the display with the most frequent lexemes beginning with ‘s’ (Figure 7).

From the s-grid, (e.g.,) typing uppercase <U> transmits “是不是” (/shìbùshì/,”yes or no”).

4. Modifications for Manipulation Challenged Users

For the subset of MC users addressed in this study, a novel **method based on wrist rotation** (**WRM**) on keyboard keys eliminates problematic knuckle-use on either keyboards or mice.

Generally, computer users (including MC users) prefer using mice with either left or right hands. For MC ***μLex*** users, the same hand continues to operate the mouse, but only for the purpose of moving the cursor on ***μLex*** grids; mouse buttons are ignored, eliminating one source of challenge. Instead, MC users specify two keyboard keys for their thumbs and little fingers (**pinkies**). The space bar and the enter key are larger than other keys, so they are likely configuration choices for right hands (Figure 5).

Figure 8: WRM Finger Thumb and Pinky Positioning

These two keys substitute for mouse buttons. To eliminate knuckle use, wrist rotation together with shorter and longer duration key-presses substitute for single and double click. So, the four actuations for WRM correspond to mouse actuations:

- Thumb Short Press (**TSP**) ~ LSC
- Thumb Long Press (**TLP**) ~ LDC
- Pinky Short Press (**PSP**) ~ RSC
- Pinky Long Press (**PLP**) ~ RDC

MC user operation of ***μLex*** applications is straightforward based on the correspondence of for mouse and WRM actuations.[4]

5. Algorithmic Generation of GridSets from Frequency Lists

μLex incorporates lexical frequency data to provide (gesture- and locatability-) optimized lexeme choices. Its basis is an algorithm that operates on lexeme frequency lists and produces intermediate data in Key/Val (i.e., 'dictionary') form. Samples of the intermediate data structure are shown in Figure 9.

μ短语 uses data from the first row of Figure 9 to produce its initial grid (Figure 6) and the 's' row to produce the grid in Figure 7.

The dictionary keys (Figure 9, first column) are short **qwerty/roman/ascii letter prefixes** (**QLP**s) which specify keyboard actuation sequences (summary cells). The dictionary values (Figure 9, second column) are mutually exclusive, frequency-ordered **lists of 30 or fewer lexemes** (**LexLists**) for the lexeme cells. QLPs can be associated with LexLists in any Romanized language or any language for which a Romanization can be made (e.g., Pinyin for Mandarin).

QLP-based LexList dictionaries are useful for defining and producing gesturally efficient UIs, where QLP lengths effectively indicate gesture counts required for lexeme input, regardless of the input device used.

The initial data for developing ***μLex*** applications is frequency lists; for English and other qwerty/roman alphabetic based languages a simple lexeme frequency list is sufficient. To provide coverage for non-lexeme input for non-qwerty languages, it may be useful to supplement the lexeme

[4] WRM technical implementation is the subject of a forthcoming paper.

frequency list with a list of single characters that appear in lexeme lists but may not be lexemes themselves. Mandarin is a good example (see Figure 10 v. Figure 11).

Key	Val
	的,了,是,在,一,和,他,不,我,有,这,人,也,上,说,着,就,地,中,对,要,你,一个,她,到,为,个,都,又,把
a	按,爱,啊,安排,按照,安全,爱情,案,暗,安,安置,案件,安伦,安慰,安哥拉 阿拉伯,爱社,安装,爱好,俺,挨,唉,爱国,爱护,按劳分配,安拉,安盟,阿,熬
ro	肉,镕基,容量,柔,荣誉
ru	如今,如下,如同,若干,乳糖,软,入党,若是,儒学,入学,乳
s	时,使,所,什么,社会,三,生活,生产,世界,事,时间,四,市场,水,岁,时候 声,谁,社会主义,思想,所以,手,上海,水平,十分,使用,少,死,受
shi	是否,事物,食,适应,实在,实际上,失去,史,食物,石油,始终,适当,诗,世纪,室 师,市委,是不是,事件,室内,实验,试,试验,氏,十几,施,失败,释迦牟尼,示
shi+	实,湿,视,世,识,势,失,士,始,拾
shib	识别
zu	坐,最后,座,作出,作品,左右,组成,最好,祖国,作者,最高,最近,嘴,做好,做到 作案,尊重,组,作家,最终,左,做出,做法,组织部,足,组合,罪犯,作风,昨天
zu+	租
zua	钻,钻石
zuc	组成部分
zug	足够
zui	嘴唇,醉,最低,最为,最初,最佳,罪,嘴里,最新
zun	遵循,遵守,尊
zuo	左权,左手,坐下,作战,昨晚
zux	祖先
zuy	足以

Figure 9: Sample of **μ短语** *Grid Specifications*

Rnk	Frq	Lex
1	75233.62	，
2	51047.79	的
3	17910.3	"
4	12763.69	了
5	11579.86	是
6	10401.01	在
202	467.15	者
203	462.16	结构
479	233.57	为什么
707	165.7	感觉
708	165.7	科学技术
1296	91.83	2．
1809	64.88	中华人民共和国
2166	52.9	①
2298	48.91	C
3257	33.94	个人所得税
3389	31.94	一九九一年
4963	19.96	１９８４年
4997	19.96	几分
4998	19.96	文化大革命
4999	19.96	东汉
5000	19.96	产卵

Figure 10: Mandarin Frequency List

Rnk	Chr	Freq
948	的	7,922,684
1	一	3,050,722
1192	是	2,615,490
86	不	2,237,915
15	了	2,128,528
346	在	2,009,181
9	人	1,867,999
347	有	1,782,004
648	我	1,690,048
247	他	1,595,761
685	肝	9,551
2457	擦	9,542
3452	簇	2,173
3100	棱	2,168
3436	檩	96
2752	荠	94
2940	麸	83
2891	秣	72
3402	蟥	61
2848	荸	55
2789	秕	39
2812	柒	17

Figure 11: Mandarin Character List

6. Measures for Productivity

For many English language computer users, qwerty keyboards are the most common **input methods** (**IME**s) for producing longer, complex, or specialized texts.

Keyboard usage leads intuitively to a GE statistic of **unicodes per gesture** (**UPG**) for producing English texts, with a resulting approximate GE for English of

$$\frac{1\ unicode}{1\ gesture} = 1.$$

Even for Mandarin, the most spoken language worldwide, qwerty keyboards are often used for producing texts. So, there are many IMEs that attempt to make Mandarin input easier and/or more efficient for users. Among these are: Microsoft Pinyin, Sogou Pinyin, Google Pinyin, Baidu Input, QQ Pinyin, and Pleco IME.

Mandarin IMEs transmit single unicodes with more than one gesture. For example, Microsoft Pinyin (Figure 12) transmits “猫” (/māo/,”cat”) with 4 keyboard gestures (**<m>**,**<a>**,**<o>**,**<2>**). Brief

examination of several commercially available Mandarin IMEs on single Mandarin unicodes indicate gesture counts in the range 2-10. Corresponding UPGs are 0.5 to 0.1.

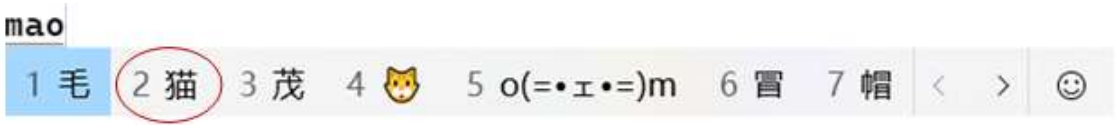

Figure 12: Using Microsoft Pinyin to transmit 猫

Although UPG is sufficient for comparing GEs for single languages, the metric is misleading for cross-linguistic comparison of ***corresponding*** texts. Consider corresponding single-lexeme texts "cat" and "猫". Since this study deals with multiple languages, it uses **lexemes per gesture** (**LPG**) statistics in addition to UPG. For example, 4 keyboard gestures (`<c>`,`<a>`,`<t>`,`<space>`) transmits "cat". So, for "cat", UPG is 1, and LPG is 0.25. For "猫" both UPG and LPG are 0.25.

In related studies, ***μLex*** application GE comparisons are via UPG and LPG summary statistics. The IMEs and associated statistics are based on frequency lists of web-based documents as sample summaries. UPG and LPG comparisons are via range statistics (min, max) and central-tendency statistics (mean, median, and mode). Additional comparisons are on lengths and elements of **Coverage Per Gesture Vectors** (**CPG**s). **CPG**s indicate the mass and accumulated mass of sample text on a per-gesture basis. For example, a **UPG CPG** of **[0.1, 0.2. 0.3, 0.4]** indicates that **10%** of the unicodes in texts can be transmitted by the sample ***μLex*** application in **1** gesture, **20%** in **2** gestures, **30%** in **3**, and **40%** in **4**.