บทที่ 3

การทดลองหาสถิติการพิมพ์ผิดในภาษาไทย

การพิมพ์ผิดในภาษาไทยอาจมีรูปแบบสอดคล้องหรือแตกต่างกับสถิติการพิมพ์ผิดในภาษาอังกฤษก็ได้ เนื่องจากในภาษาไทยนั้นมีข้อแตกต่างจากภาษาอังกฤษคือ อักขระไทยนั้นมีหลายระดับ เช่น วรรณยุกต์ สระบน สระล่าง เพิ่มเติมขึ้นมา ดังนั้นในงานวิจัยนี้จะเน้นทำการวิเคราะห์สถิติใน 2 ประเด็น คือ 1.หาสัดส่วนความผิดพลาดในการพิมพ์ 4 ชนิด (เกิน ตก แทนที่ สลับ) จากคำที่พิมพ์ผิดทั้งหมด 2.หาตำแหน่งอักขระต้นเหตุของคำผิด ซึ่งได้แนวคิดมาจากหลายๆ งานวิจัยซึ่งได้สังเกตพฤติกรรมการพิมพ์ผิดในภาษาอังกฤษไว้ ดังต่อไปนี้

งานวิจัยของ Damerau [10] ได้ค้นพบว่า ประมาณ 80 เปอร์เซ็นต์ของคำที่พิมพ์ผิดทั้งหมดเกิดขึ้นจากการประกอบกันของความผิดพลาด 4 ชนิด คือ 1.การแทรก (Insertion) 2.การลบ (Deletion) 3.การแทนที่ (Substitution) 4.การสลับ (Transposition)

นอกจากนี้มีผลการวิจัยหลายงานที่ได้สังเกตพฤติกรรมในการพิมพ์พบว่า คนมักจะพิมพ์ผิดที่ตำแหน่งอักขระตัวแรกสุดของคำน้อยกว่าตัวอักขระตำแหน่งอื่นๆ เช่น ในงานวิจัยของ Pollock and Zamora [15] พบว่ามีคำผิดที่เกิดจากอักขระแรกเพียง 3.3 เปอร์เซ็นต์ และจากงานวิจัยของ Kukich [9] พบเพียง 15 เปอร์เซ็นต์ โดยที่ทั้งสองงานใช้ความยาวเฉลี่ยของคำที่มี 3 ถึง 4 ตัวอักษรเท่านั้นในการทดสอบ และในงานวิจัยของ Kukich [9] ได้สรุปไว้ว่ามีเพียงส่วนน้อยมากของความผิดพลาดที่เกิดจากอักขระตัวแรกของคำ

ในบทนี้แบ่งเนื้อหาออกเป็น 2 ส่วนคือ การออกแบบการทดลอง และการวิเคราะห์หาสถิติ

3.1 การออกแบบการทดลองหาสถิติการพิมพ์ผิดในภาษาไทย

ในการออกแบบการทดลองได้เลือกข้อความที่มีลักษณะเป็นบทสนทนามาใช้ในการทดสอบการพิมพ์เนื่องจากต้องการนำไปใช้กับหุ่นยนต์สนทนา ซึ่งได้มาจากหนังสือทางช้างเผือก [16] ที่เป็นวรรณกรรมเยาวชน แล้วนำมาดัดแปลงบางส่วน คือ ได้ตัดมาเฉพาะส่วนที่เป็นบทสนทนาและเปลี่ยนชื่อตัวละครบางตัวให้เป็นคำที่มีความหมายในพจนานุกรม เพื่อที่จะได้ไม่ถูกเข้าใจผิดมองว่าเป็นคำผิดในการที่ต้องใช้วิเคราะห์ต่อไป โดยที่แบบทดสอบนั้นมีจำนวนทั้งสิ้น 4,391 คำ หรือ 15,509 อักขระ (โดยไม่นับอักขระว่าง) ซึ่งบทความที่ใช้ทดสอบนี้ได้แสดงไว้ในภาคผนวก ก

การทดลองเริ่มจากหาอาสาสมัครจำนวน 30 คน ประกอบด้วยผู้ที่มีทักษะในการพิมพ์ 15 คน และไม่มีทักษะในการพิมพ์ 15 คน มาพิมพ์แบบทดสอบเดียวกัน ซึ่งใช้โปรแกรมเอดิทพลัสเวอร์ชัน 2.11 ให้อาสาสมัครใช้ในการพิมพ์แล้วบันทึกไว้ในรูปแบบไฟล์ข้อความ (Text File) และใน

การทดลองได้แบ่งแบบทดสอบออกเป็น 3 ชุดย่อย ให้ผู้ทดสอบทำทีละชุด เพื่อลดปัจจัยด้านความอ่อนล้าซึ่งอาจเป็นสาเหตุของประสิทธิภาพในการพิมพ์ แล้วบันทึกเวลาในการทำแบบทดสอบของแต่ละชุด แต่สุดท้ายแล้วก็จะนำผลการทดสอบพิมพ์ทั้ง 3 ชุดของแต่ละคนมารวมเป็นชุดเดียวกันเพื่อรอทำการวิเคราะห์ต่อไป

การควบคุมการทดลองให้มีสภาพแวดล้อมเหมือนกัน ทำโดยให้อาสาสมัครมานั่งพิมพ์ที่ห้องปฏิบัติการของผู้วิจัย (ซึ่งมีสภาพแวดล้อมเหมือนสำนักงานทั่วไป) โดยมีผู้วิจัยนั่งควบคุมดูแลอยู่ข้างๆ โดยทำความเข้าใจกับผู้ทดสอบก่อนว่าให้พิมพ์เสมือนกับว่าเป็นการสนทนากับเพื่อนผ่านทางอินเทอร์เน็ต คือไม่เร็วหรือช้าจนเกินไป และถ้าหากพิมพ์ผิดไม่ต้องแก้ไขก็ได้ให้พิมพ์ต่อไปเลย แต่ถ้าหากเคยมือมักจะแก้ไขก็ปล่อยไปตามธรรมชาติของบุคคลนั้น นอกจากนี้ยังสอบถามก่อนทำการพิมพ์ถึงทักษะของอาสาสมัครแต่ละคนว่ามีทักษะในการพิมพ์หรือไม่ โดยทักษะในการพิมพ์นั้นนิยามตามงานวิจัยของ Grudin [11] ที่นิยามไว้ในการทดลองของเขาว่า

- **ผู้มีทักษะในการพิมพ์** (Skilled Typist) คือ ผู้พิมพ์สัมผัสได้ หรือเคยเรียนพิมพ์ดีด
- **ผู้ไม่มีทักษะในการพิมพ์** (Novice Typist) คือ ผู้ที่ไม่สามารถพิมพ์สัมผัสได้

งานวิจัยของ Grudin [11] ได้กล่าวถึงการศึกษารูปแบบของการพิมพ์ผิดในระดับความชำนาญต่างๆ กัน ได้ผลคือ ผู้ที่มีทักษะมักพิมพ์เกินใน 2 คีย์ที่อยู่ติดกัน ในขณะที่ผู้ไม่มีทักษะมักผิดพลาดจากการพิมพ์ผิด และการพิมพ์ผิดนั้นมักเกิดจากคีย์ที่อยู่ติดกัน *แต่ไม่ได้บอกรายละเอียดวิธีการทดลองและสภาพแวดล้อมในการทดลองไว้*

การทดลองของ Grudin บอกเพียงแต่นิยามของการมีทักษะในการพิมพ์ ซึ่งได้ทำการทดลองกับผู้มีทักษะ 6 คน และไม่มีทักษะ 8 คน โดยให้พิมพ์ข้อความจากนิตยสารคนละ 60,000 ตัวอักษร แล้วพบว่าผู้ที่มีทักษะพิมพ์ผิดพลาดเฉลี่ยคิดเป็น 1.9 เปอร์เซ็นต์ และผู้ที่ไม่มีทักษะพิมพ์ผิดพลาดเฉลี่ยคิดเป็น 3.2 เปอร์เซ็นต์ นอกจากนี้ยังพบว่าคำผิดที่เกิดจากลักษณะการแทนที่ทั้งสิ้นประมาณ 3,800 คำนั้น มีผลมาจากการแทนที่ตัวอักษรปุ่มที่อยู่ข้างเคียงลงไปคิดเป็นประมาณ 58 เปอร์เซ็นต์

งานวิจัยของ Damerau [10] และ Pollock and Zamora [15] นั้นไม่ได้บอกถึงรายละเอียดและวิธีการในการเก็บรวบรวมคำผิดที่จะนำมาใช้ในการวิเคราะห์ บอกเพียงแต่ว่านำข้อความที่ใช้ทดสอบมาจากฐานข้อมูลความรู้ทางวิทยาศาสตร์เป็นจำนวน 170,016 คำ และ 25,000,000 คำ ตามลำดับงานวิจัย สังเกตได้ว่าเป็นการรวบรวมสถิติมาจากข้อความที่พิมพ์ไว้แล้ว ไม่ได้มีการควบคุมสภาพแวดล้อมในการเก็บรวบรวมข้อมูล

ในงานวิจัยนี้จึงออกแบบวิธีและการควบคุมสภาพแวดล้อมขึ้นมาเอง โดยอาศัยข้อมูลที่มีให้คือ การนิยามเรื่องทักษะของผู้พิมพ์และปริมาณคำที่ใช้ในการทดลอง ส่วนองค์ประกอบนอกเหนือจากนั้นผู้วิจัยได้ทำการกำหนดและควบคุมให้เป็นรูปแบบเดียวกันตลอดการทดลอง

โดยนึกถึงวัตถุประสงค์ที่จะนำไปใช้กับหุ่นยนต์สนทนา จึงเลือกข้อความที่เป็นบทสนทนามาเป็นแบบทดสอบในการเก็บสถิติคำผิด

3.2 การวิเคราะห์ผลการทดลอง

เมื่อได้ผลทดสอบการพิมพ์ที่ได้จากแบบทดสอบดังที่กล่าวไว้ในหัวข้อ 3.1 แล้ว ขั้นตอนการทำงานต่อไปคือ การแยกคำที่ไม่เหมือนต้นฉบับออกมาจากผลการทดสอบที่ได้จากการพิมพ์ โดยบันทึกคำต่างทั้งหมดออกมาและบันทึกคำเดิมไว้คู่กัน เพื่อรอทำการวิเคราะห์ต่อไป ในงานวิจัยนี้ใช้โปรแกรม Beyond Compare เวอร์ชัน 2.2.3 ช่วยในการเปรียบเทียบหาคำที่ต่างไปจากต้นฉบับ

คำต่างทั้งหมดที่ได้จากโปรแกรมช่วยเปรียบเทียบนั้นจะถูก*นำมาทำการกรองด้วยมืออีกที โดยตัดกรณีพิมพ์ตกหรือพิมพ์เกินมาเป็นคำ*ออกไป ไม่คิดรวมลงไปในสถิติ

การวิเคราะห์จะแบ่งออกเป็น 2 ประเด็นคือ หาสัดส่วนความผิดพลาดในการพิมพ์ทั้ง 4 ชนิด และหาตำแหน่งอักขระต้นเหตุของคำผิด ดังต่อไปนี้

3.2.1 การหาสัดส่วนความผิดพลาดในการพิมพ์

การวิเคราะห์หาสัดส่วนของสาเหตุต่างๆ ที่ทำให้เกิดคำผิด โดยมีขั้นตอนในการแบ่งแยกคำที่ไม่เหมือนต้นฉบับออกเป็นกลุ่มต่างๆ ตามสาเหตุที่ทำให้เกิดความผิดพลาดดังนี้

1. เทียบผลลัพธ์การพิมพ์เทียบกับต้นฉบับที่ใช้พิมพ์ โดยคำที่ไม่เหมือนต้นฉบับให้ถือว่าเป็นคำต่าง

2. นำคำต่างทั้งหมดที่ได้มาเทียบกับคำในพจนานุกรมภาษาไทยจากราชบัณฑิตยสถาน [17] มาเป็นเกณฑ์ต่อไปในการแยกคำต่างออกเป็น 2 กลุ่มใหญ่คือ คำถูก และคำผิด โดยคำที่ไม่พบในพจนานุกรมให้ถือว่าเป็นคำผิด

3. นำคำผิดที่ไม่พบในพจนานุกรมมาแยกออกเป็นกลุ่มย่อยโดยใช้ระยะแก้ไข [10] มาช่วยแบ่งแยกคำผิดออกเป็น 2 กลุ่มคือ ระยะแก้ไขเท่ากับ 1 และระยะแก้ไขมากกว่า 1

4. นำคำผิดที่มีระยะแก้ไขเท่ากับ 1 ทั้งหมดมาแยกวิเคราะห์หาสัดส่วนของความผิดพลาดในการพิมพ์ทั้ง 4 ชนิด [10] คือ การพิมพ์เกิน ตก แทนที่ สลับ มาเป็นกลุ่มหลักในการแบ่ง

- นอกจากนี้ยังมีความผิดพลาดจากการจงใจของผู้ทดสอบ เช่น ผู้ทดสอบสะกดผิดเองโดยเกิดจากทักษะการใช้ภาษาของผู้ทดสอบ เพิ่มขึ้นมาอีก 1 กลุ่ม
- การพิมพ์เกินนั้นอาจแบ่งออกได้เป็น 2 แบบคือ ไม่ซ้ำ และซ้ำตัวก่อนหน้า (Repeat Key)

จากขั้นตอนดังกล่าวทำให้สามารถแบ่งคำทั้งหมดออกได้เป็น 10 กลุ่มย่อย ดังตารางที่ 3.1

ตารางที่ 3.1 สรุปจำนวนความผิดพลาดแบบต่างๆ ของการพิมพ์

id	tw	2w	syn	d>1	int	s	all	d	r	Total err	skill	time	%err
3	37	0	6	1	1	39	22	16	16	11	149	0	106.44	3.39
8	19	3	20	0	3	4	4	6	11	1	71	0	112.04	1.62
9	41	3	13	6	2	47	1	7	35	2	157	0	100.15	3.58
11	20	1	34	1	2	22	15	11	12	5	123	0	138.31	2.80
13	36	3	3	0	1	44	7	6	9	7	116	0	101.98	2.64
14	58	4	21	3	2	77	1	11	24	4	205	0	122.65	4.67
15	9	3	19	1	3	19	5	2	8	1	70	0	131.78	1.59
17	23	11	3	20	0	44	12	110	4	5	232	0	81.73	5.28
18	20	0	4	1	2	15	2	6	5	0	55	0	119.53	1.25
20	16	2	6	1	0	52	3	8	3	0	91	0	112.68	2.07
24	16	3	31	0	0	17	14	6	6	5	98	0	149.76	2.23
25	52	6	35	23	6	168	13	41	11	24	379	0	181	8.63
26	1	1	0	0	0	1	0	0	1	0	4	0	218	0.09
27	22	1	19	0	2	24	10	4	3	8	93	0	107.12	2.12
28	28	3	32	3	2	19	6	10	5	2	110	0	133	2.51
													127.74	2.96

ก

id	tw	2w	syn	d>1	int	s	all	d	r	Total err	skill	time	%err
1	56	7	2	5	1	25	2	24	85	9	216	1	65.82	4.92
2	86	1	2	1	2	85	11	30	42	8	268	1	67.87	6.10
4	110	0	32	3	1	117	20	76	20	8	387	1	138.98	8.81
5	77	8	7	37	5	104	9	72	16	10	345	1	137.9	7.86
6	76	17	3	15	1	145	14	126	10	27	434	1	116.13	9.88
7	12	8	4	7	0	33	5	46	19	16	150	1	57	3.42
10	39	7	14	6	0	52	10	25	48	7	208	1	90	4.74
12	78	18	5	32	2	84	25	121	29	53	447	1	106.35	10.18
16	47	7	24	10	3	40	16	51	30	19	247	1	95.81	5.63
19	38	3	3	12	1	48	20	33	15	5	178	1	98.2	4.05
21	17	4	2	2	1	11	6	25	8	6	82	1	55	1.87
22	57	2	36	8	5	65	4	18	20	13	228	1	98	5.19
23	39	4	2	4	2	40	13	16	7	2	129	1	130.94	2.94
29	16	1	12	6	1	12	3	6	8	2	67	1	64	1.53
30	44	3	30	2	0	21	8	47	25	26	206	1	60	4.69
													92.13	5.45

ข

ตารางที่ 3.1 ก แสดงถึงความผิดพลาดในการพิมพ์ของผู้ไม่มีทักษะ มีค่าเฉลี่ยของการพิมพ์ผิดเป็น 2.96 เปอร์เซ็นต์ (โดยค่าเฉลี่ยคิดจากเปอร์เซ็นต์ความผิดพลาดของแต่ละคนในกลุ่มมารวมกันหารด้วยจำนวนคนในกลุ่ม ซึ่งในที่นี้คือ 15) นอกจากนี้จะเห็นได้ว่าความผิดพลาดเฉลี่ยในการพิมพ์เรียงลำดับความสำคัญตามปริมาณที่พบได้ดังนี้คือ แทนที่>เกิน>ตก>สลับ

ตารางที่ 3.1 ข แสดงถึงความผิดพลาดในการพิมพ์ของผู้มีทักษะ มีค่าเฉลี่ยของการพิมพ์ผิดเป็น 5.45 เปอร์เซ็นต์ จะเห็นได้ว่าความผิดพลาดเฉลี่ยในการพิมพ์เรียงลำดับความสำคัญตามปริมาณที่พบได้ดังนี้คือ แทนที่>เกิน>ตก>สลับ

สรุปได้ว่าผู้ที่ไม่มีทักษะจะพิมพ์ผิดน้อยกว่า ทั้งนี้คาดว่าเป็นเพราะวิธีการพิมพ์ของผู้ไม่มีทักษะนั้นจะใช้ตามองแป้นพิมพ์เกือบตลอดเวลา ซึ่งทำให้ใช้เวลาในการพิมพ์มากกว่า (ผู้ไม่มีทักษะใช้เวลาเฉลี่ยคนละ 127.74 นาที แต่ผู้มีทักษะใช้เวลา 92.13 นาที) ส่วนในด้านปริมาณความผิดพลาดของแต่ละสาเหตุก็เป็นไปในทิศทางเดียวกัน คือ ทั้ง 2 กลุ่มพิมพ์ผิดแบบแทนที่มากที่สุด ตามมาด้วยพิมพ์เกิน พิมพ์ตก และพิมพ์สลับ ตามลำดับ

ความผิดพลาดในการพิมพ์ทั้งหมดของผู้มีทักษะและไม่มีทักษะเฉลี่ยคิดเป็น 4.21 เปอร์เซ็นต์ จากคำที่ใช้ในแบบทดสอบการพิมพ์ 4,391 คำ กับอาสาสมัครทำแบบทดสอบจำนวน 30 คน โดยใช้แบบทดสอบเดียวกันในการทดสอบ ซึ่งคำอธิบายของตัวย่อที่ใช้ในตารางมีดังนี้

1w : ผิดจากต้นฉบับ กลายเป็นคำถูก 1 คำ เช่น มาก ไปเป็น มาหก

2w : ผิดจากต้นฉบับ กลายเป็นคำถูก 2 คำ เช่น มาก ไปเป็น มาด

syn : ผิดจากต้นฉบับไปเป็นคำที่มีความหมายคล้ายกัน เช่น จ้ะ ไปเป็น จ๊ะ

d>1 : พิมพ์ผิดมากกว่า 1 อักขระ

int : ไม่พบในพจนานุกรม แต่เกิดจากความจงใจของผู้ใช้ เช่น พราก เป็น พลาก

s : การพิมพ์แทนที่

ii : การพิมพ์เกินแบบซ้ำตัวก่อนหน้า (Repeat Key)

i : การพิมพ์เกินแบบไม่ซ้ำ

d : การพิมพ์ตก

r : การพิมพ์สลับที่

จากคำอธิบายตัวย่อที่ใช้ในตารางที่ 3.1 ซึ่งแบ่งแยกความผิดพลาดแบบต่างๆ ของการพิมพ์ออกเป็น 10 กลุ่มย่อยนั้น ความผิดพลาด 3 กลุ่มแรกเป็นคำต่างจากต้นฉบับแต่กลายไปเป็นคำถูกที่มีความหมายในพจนานุกรม ต่อมากลุ่มที่ 4-10 ถือว่าเป็นคำผิดทั้งหมด เพราะไม่พบในพจนานุกรม และกลุ่มที่ 6-10 เป็นกลุ่มที่เราสนใจเป็นพิเศษ ซึ่งจะนำมาวิเคราะห์สัดส่วนอย่างละเอียดในรูปแบบของเปอร์เซ็นต์ต่อไป โดยจะเทียบแต่ละกรณีที่สนใจกับจำนวนคำผิดทั้งหมด ได้ดังตารางที่ 3.2 ดังนี้

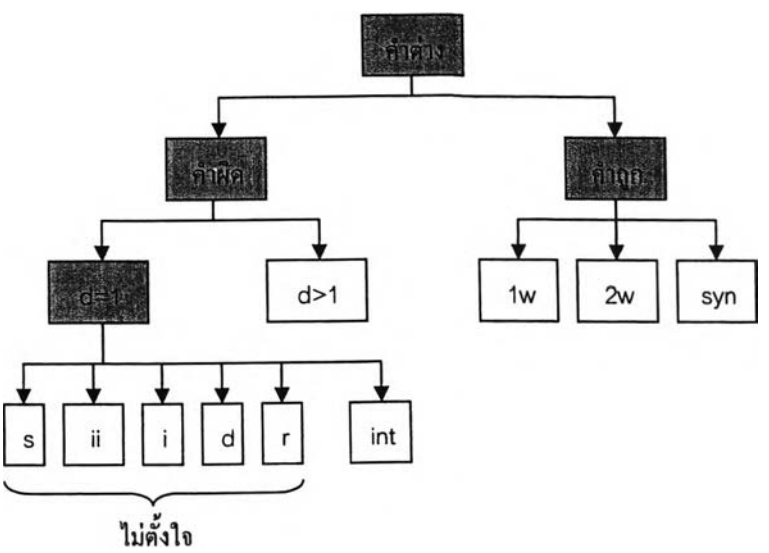
ตารางที่ 3.2 สรุปสัดส่วนความผิดพลาดของคำผิด

สาเหตุความผิดพลาด	ตัวย่อ	เปอร์เซ็นต์ความผิดพลาด	รวม
ระยะแก้ไขมากกว่า 1	d>1	4.41	6.46
จงใจผิด	int	2.06	
พิมพ์แทนที่	s	39.51	93.54
พิมพ์เกินแบบซ้ำอักขระก่อนหน้า	ii	9.08	
พิมพ์เกินแบบไม่ซ้ำ	i	20.82	
พิมพ์ตก	d	17.43	
พิมพ์สลับ	r	6.70	

จากตารางที่ 3.2 จะเห็นได้ว่า *การทดลองกับภาษาไทยนั้นคำผิดส่วนใหญ่เกิดจากความผิดพลาดทั้ง 4 กรณีประกอบกันซึ่งคิดเป็น 93.54 เปอร์เซ็นต์จากคำผิดที่พบทั้งหมด* ซึ่งสอดคล้องตามงานวิจัยของ Damerau [10] ที่ได้ค้นพบว่าในภาษาอังกฤษจะพบประมาณ 80 เปอร์เซ็นต์ของคำที่พิมพ์ผิดทั้งหมด

นอกจากนี้ยังพบว่าความผิดพลาดทั้ง 4 กรณีนั้น สามารถนำเรียงลำดับความสำคัญตามปริมาณที่พบได้ดังนี้คือ แทนที่>เกิน>ตก>สลับ

จากตารางที่ 3.1 สามารถนำมาเขียนเป็นรูปแสดงโครงสร้างของคำต่างที่สามารถแบ่งออกเป็น 10 กลุ่มย่อย ได้ดังรูป 3.1



รูปที่ 3.1 การแบ่งกลุ่มของคำที่พิมพ์ต่างไปจากต้นฉบับ

จากรูปที่ 3.1 กลุ่มที่ไม่ได้แรเงา คือ กลุ่มย่อยทั้ง 10 ชนิด และใช้คำอธิบายตัวย่อเช่นเดียวกันกับตารางที่ 3.1

3.2.2 การหาตำแหน่งอักขระต้นเหตุของคำผิด

ขั้นตอนการหาตำแหน่งอักขระต้นเหตุของคำผิดนี้ จะนำเฉพาะคำผิดแบบไม่ตั้งใจที่มีระยะแก้ไขเท่ากับ 1 เท่านั้นมาทำการวิเคราะห์ ซึ่งประกอบไปด้วยความผิดพลาด 4 ชนิด (พิมพ์เกิน ตกแทนที่ สลับ) มาหารูปแบบว่ามักจะพิมพ์ผิดที่อักขระใดเป็นส่วนใหญ่ หรือกล่าวได้ว่ามักพิมพ์ผิดที่ตำแหน่งกลางคำหรือท้ายคำมากกว่ากัน เพื่อที่จะได้นำไปออกแบบอัลกอริทึมในการแก้ไขคำผิดต่อไป

มีสูตรในการคำนวณหาตำแหน่งอักขระต้นเหตุของคำผิด ดังนี้

$$(2P-1) / 2L \qquad (3.1)$$

โดยที่ P คือตำแหน่งที่ผิดพลาด และ L คือความยาวของคำ สามารถแสดงการคำนวณตามสูตรได้ดังตัวอย่างที่ 3.1-3.3

ตัวอย่างที่ 3.1 แสดงกรณีพิมพ์ผิดจากการแทนที่ เช่น คำว่า มาก พิมพ์ผิดเป็น มสก

จากตัวอย่างคำผิดนี้ จะได้ค่าตำแหน่งที่ผิดพลาดเป็น 2 และค่าความยาวของคำเป็น 3 นำไปแทนในสมการที่ 3.1 ได้เป็น

ตำแหน่งอักขระต้นเหตุของคำผิด = ((2*2) -1) / (2*3)
= (4-1) / 6
= 3 / 6
= 1 / 2
= 0.5

ดังนั้น ตำแหน่งอักขระต้นเหตุของคำผิดนี้ = 50 เปอร์เซ็นต์ของความยาวคำ

ตัวอย่างที่ 3.2 แสดงกรณีพิมพ์ผิดจากการพิมพ์เกิน เช่น คำว่า เดียว พิมพ์ผิดเป็น เดียนว

จากตัวอย่างคำผิดนี้ จะได้ค่าตำแหน่งที่ผิดพลาดเป็น 5 และค่าความยาวของคำเป็น 6 นำไปแทนในสมการที่ 3.1 ได้เป็น

ตำแหน่งอักขระต้นเหตุของคำผิด = ((2*5) -1) / (2*6)
= (10-1) / 12
= 9 / 12

= 3 / 4

= 0.75

ดังนั้น ตำแหน่งอักขระต้นเหตุของคำผิดนี้ = 75 เปอร์เซ็นต์ของความยาวคำ

ตัวอย่างที่ 3.3 แสดงกรณีพิมพ์ผิดจากการพิมพ์สลับ เช่น คำว่า แขน พิมพ์ผิดเป็น แนข

จากตัวอย่างคำผิดนี้ จะได้ค่าตำแหน่งที่ผิดพลาดเป็น 2.5 (เนื่องจากเป็นความผิดพลาดจากการพิมพ์สลับที่) และค่าความยาวของคำเป็น 3 นำไปแทนในสมการที่ 3.1 ได้เป็น

ตำแหน่งอักขระต้นเหตุของคำผิด = ((2*2.5) -1) / (2*3)

= (5-1) / 6

= 4 / 6

= 2 / 3

= 0.6667

ดังนั้น ตำแหน่งอักขระต้นเหตุของคำผิดนี้ = 66.67 เปอร์เซ็นต์ของความยาวคำ

ทำการคำนวณเช่นนี้กับคำผิดทุกๆ คำ แล้วนำมาหาค่าเฉลี่ยแล้วคิดให้อยู่ในรูปแบบของเปอร์เซ็นต์ ซึ่งสรุปได้ดังตารางที่ 3.3

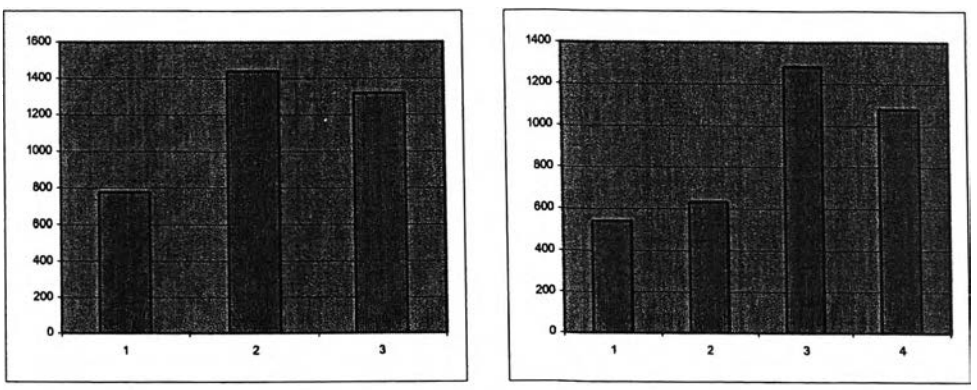
ตารางที่ 3.3 สรุปตำแหน่งอักขระต้นเหตุของคำผิด

สาเหตุความผิดพลาด	ตัวย่อ	เปอร์เซ็นต์เฉลี่ยของตำแหน่ง	ค่าเบี่ยงเบนมาตรฐาน
พิมพ์แทนที่	s	53.96	26.11
พิมพ์เกินแบบซ้ำอักขระก่อนหน้า	ii	52.39	18.74
พิมพ์เกินแบบไม่ซ้ำ	i	69.58	24.48
พิมพ์ตก	d	53.83	26.54
พิมพ์สลับ	r	57.71	18.50

จากตารางที่ 3.3 จะเห็นได้ว่า *ไม่ว่าจะเป็นกรณีความผิดพลาดใดๆ ตำแหน่งอักขระที่ผิดนั้นมักจะอยู่ตรงกลางแต่ค่อนไปทางด้านหลังเล็กน้อย คิดเป็นค่าเฉลี่ยของทุกกรณีได้เป็นตำแหน่งที่ 58.36 เปอร์เซ็นต์ของความยาวคำ* ซึ่งไม่ขัดกับผลการวิจัยการพิมพ์ภาษาอังกฤษของ Kukich [9] ที่ได้สรุปไว้ว่ามีเพียงส่วนน้อยมากของความผิดพลาดที่เกิดจากอักขระตัวแรกของคำ

ความยาวเฉลี่ยที่ใช้ในการทดสอบ สามารถหาได้จาก จำนวนอักขระหารด้วยจำนวนคำ โดยที่แบบทดสอบนั้นมีจำนวนทั้งสิ้น 4,391 คำ หรือ 15,509 อักขระ (โดยไม่นับอักขระว่าง) จะมีความยาวเฉลี่ยของคำเป็น 15,509/4,391 คิดเป็นความยาวเฉลี่ยเท่ากับ 3.53 อักขระต่อคำ

จากข้อมูลในตารางที่ 3.3 หากนำมาวาดรูปการแจกแจง (Distribution) จะได้รูปออกมาเป็นการแจกแจงปกติต่อกันหลายรูปทำให้ยากต่อการวิเคราะห์ด้วยรูป แต่ถ้าหากวาดเป็นรูปฮิสโทแกรม (Histogram) ก็จะเห็นได้ไม่ชัดเจน จึงนำข้อมูลออกมาแสดงในรูปของแผนภูมิแท่งโดยแบ่งข้อมูลทั้งหมดที่ได้จากการพิมพ์ผิดทุกกรณีด้วยค่าเปอร์เซ็นต์ตำแหน่งอักขระที่ผิด ซึ่งจะแบ่งข้อมูลออกเป็นกลุ่มตามความยาวของคำ ตั้งแต่ความยาว 3-4 ตัวอักษร (เนื่องจากความยาวเฉลี่ยของคำที่ใช้ในการเก็บสถิตินี้มีค่าเฉลี่ยที่ 3.53 อักขระต่อคำ) สามารถแสดงได้ดังรูปที่ 3.2



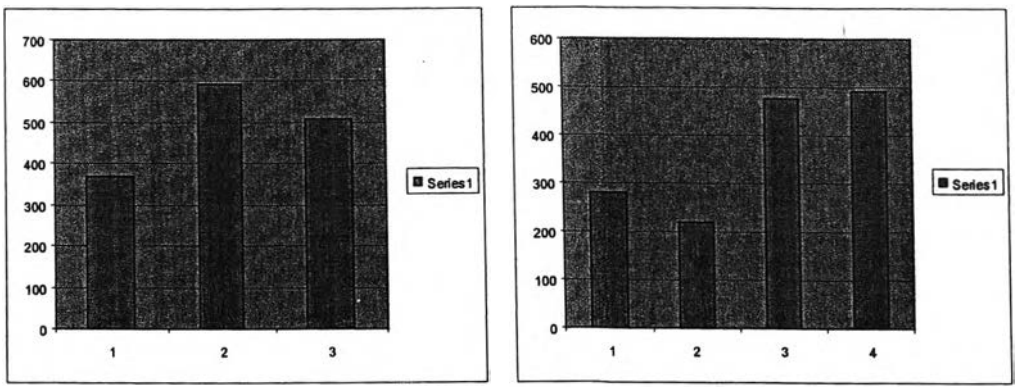
ก ข

รูปที่ 3.2 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบ่งตามความยาวคำ

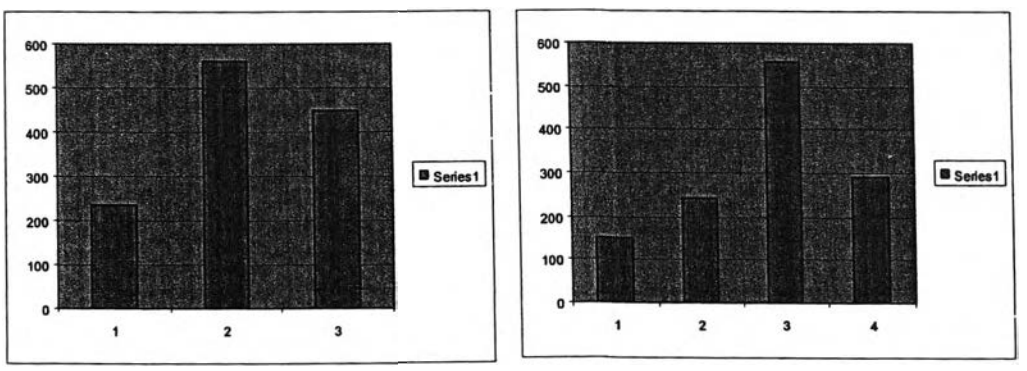
จากรูปที่ 3.2 แสดงให้เห็นถึงปริมาณความหนาแน่นของตำแหน่งอักขระต้นเหตุที่ผิด ซึ่งรูป ก ข แสดงถึงการแบ่งข้อมูลออกเป็นช่วงเท่าๆ กัน ตามความยาวของจำนวนอักขระในคำ (ซึ่งมีค่าเป็น 3-4 ตามลำดับ) จะเห็นได้ว่าไม่ว่าจะเป็นความยาวใดๆ ตั้งแต่ 3-4 นั้นจะมีความหนาแน่นอยู่ที่กลุ่มตำแหน่งตรงกลาง แต่ค่อนข้างไปทางด้านหลังเล็กน้อย สอดคล้องกับตารางที่ 3.3 ซึ่งสรุปไว้ว่า *ไม่ว่าจะเป็นกรณีความผิดพลาดใดๆ ตำแหน่งอักขระที่ผิดนั้นมักจะอยู่ตรงกลางแต่ค่อนไปทางด้านหลังเล็กน้อย คิดเป็นค่าเฉลี่ยของทุกกรณีได้เป็นตำแหน่งที่ 58.36 เปอร์เซ็นต์ของความยาวคำ*

การแบ่งข้อมูลออกเป็นช่วงที่เท่าๆ กัน เริ่มจากนำค่า 100 (ซึ่งเป็นค่าสูงสุดของเปอร์เซ็นต์) หารด้วยความยาวของคำ เช่น ต้องการวาดรูปความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดที่มีความยาวเป็น 3 ก็จะได้ช่วงข้อมูล 3 ช่วง โดยมีความกว้างของแต่ละช่วงข้อมูลเป็น 0-33.33, 33.34-66.66 และ 66.67-100 ตามลำดับ เมื่อแบ่งข้อมูลออกเป็นกลุ่มแล้วก็จะทำการนับความถี่ของแต่ละกลุ่ม แล้วนำไปสร้างรูปแผนภูมิแท่งโดยแกนตั้งแสดงความถี่ของจำนวนข้อมูลแต่ละชุด และแกนนอนแสดงถึงช่วงข้อมูลแต่ละกลุ่ม (ดังในรูป 3.2 ก ช่วงข้อมูลกลุ่มที่ 1 คือข้อมูลที่มีค่าตำแหน่งอักขระต้นเหตุคำผิดเป็น 0-33.33 เปอร์เซ็นต์ และเลข 2 ในแกนนอนแสดงถึงช่วงข้อมูลกลุ่มที่มีค่าตำแหน่งอักขระต้นเหตุคำผิดเป็น 33.34-66.66 เปอร์เซ็นต์ กลุ่มสุดท้ายคือเลข 3 ใน

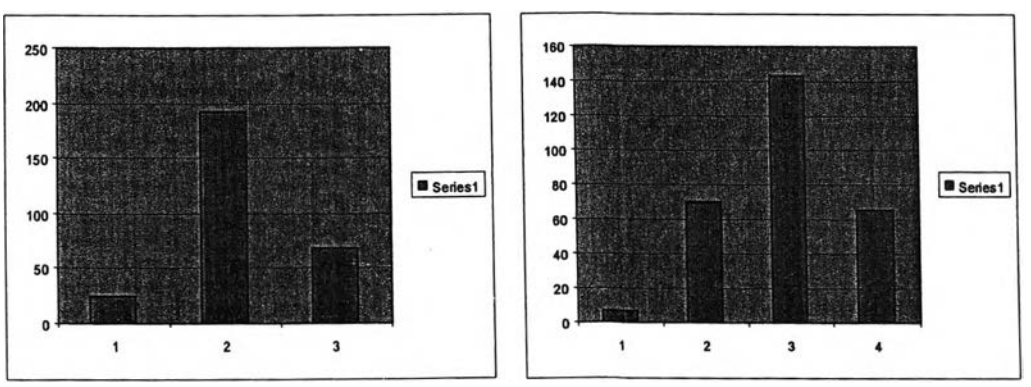
แกนนอนแสดงถึงช่วงข้อมูลกลุ่มที่มีค่าตำแหน่งอักขระต้นเหตุคำผิดเป็น 66.67-100 เปอร์เซ็นต์) ซึ่งการวาดรูปแบบนี้จะง่ายต่อการออกแบบลำดับตำแหน่งของอัลกอริทึมการแก้ไขคำผิดในภายหลัง และหากต้องการดูกราฟของแต่ละกรณีที่พิมพ์ผิดสามารถดูได้จากรูปที่ 3.3-3.6



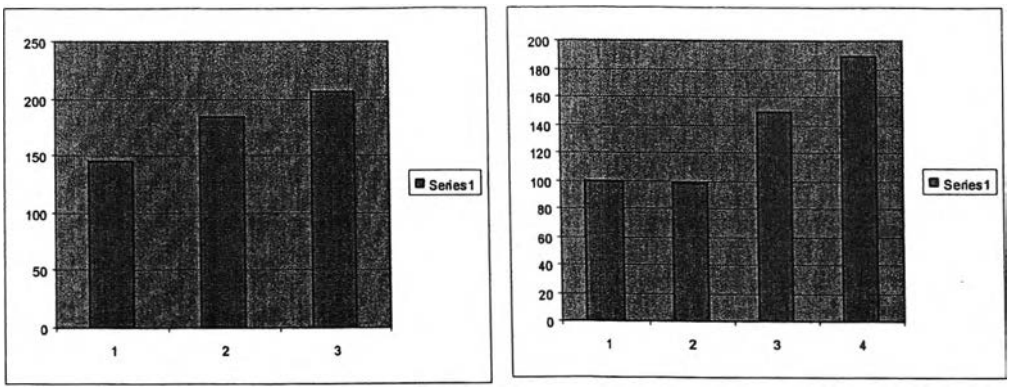
รูปที่ 3.3 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์แทนที่



รูปที่ 3.4 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์เกิน



รูปที่ 3.5 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์สลับ



รูปที่ 3.6 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์ตก