บทที่ 4
อักขระแบ่งคำ

4.1 ความหมายของอักขระแบ่งคำ

อักขระแบ่งคำ คือตัวอักษรพิเศษที่ใช้แบ่งแยกขอบเขตของคำแต่ละคำในข้อความ สำหรับในข้อความภาษาอังกฤษ อักขระแบ่งคำนี้ ก็คือเว้นวรรค (space) ซึ่งแบ่งคำในภาษาอังกฤษออกจากกัน รวมทั้งเครื่องหมายวรรคตอนอื่นๆด้วย เช่น , ; " เป็นต้น

สำหรับในข้อความภาษาไทย เมื่อผ่านมอดูลการตัดคำ จะต้องมีอักขระแบ่งคำแทรกระหว่างคำแต่ละคำเพื่อให้สามารถแบ่งแยกคำได้อย่างถูกต้องและง่ายดาย

4.2 ชนิดของอักขระแบ่งคำ

อักขระแบ่งคำที่ใช้สำหรับข้อความภาษาไทยนั้น เราอาจแบ่งออกได้เป็น 3 ประเภท คือ

1. ตัวอักษรที่พิมพ์ไม่ได้ (non-printable character) ได้แก่ รหัสควบคุม (control code) ต่างๆ คือ รหัสตั้งแต่ 00_{16} ถึง $1F_{16}$ ในตารางรหัสแอสกี

2. ตัวอักษรที่เป็นอักขระแบ่งคำโดยธรรมชาติ ได้แก่ อักษรที่เป็นอักขระแบ่งคำในภาษาอังกฤษ

3. อักษรปกติอื่นๆ

4.3 ปัญหาของการแทรกอักขระแบ่งคำ

ถ้าอักขระแบ่งคำที่แทรกลงไปในข้อความภาษาไทยเป็นตัวอักษรตัวหนึ่งเช่นเดียวกับตัวอักษรที่เป็นข้อมูลจริงซึ่งมีอยู่เดิม การที่จะแยกความแตกต่างของอักษรที่เป็นอักขระแบ่งคำกับอักษรที่เป็นข้อมูลจริงให้ได้ จึงต้องอาศัยขั้นตอนวิธีที่เหมาะสม ในการแทรกอักขระแบ่งคำ

และการตรวจรู้อักขระแบ่งคำ

4.4 การเลือกอักขระแบ่งคำ

ดังได้กล่าวมาแล้วว่า ประเภทของตัวอักษรที่เป็นอักขระแบ่งคำมีหลายชนิด ซึ่งความเหมาะสมของแต่ละชนิดขึ้นอยู่กับงานแต่ละงาน และความต้องการของผู้ใช้ อย่างไรก็ตาม เราสามารถกำหนดคุณสมบัติที่เหมาะสมของตัวอักษรที่เป็นอักขระแบ่งคำ ได้ดังนี้

1. ควรจะเป็นตัวอักษรที่มีความถี่ต่ำ คือมีโอกาสพบน้อยในข้อความโดยทั่วๆไป ทั้งนี้ เพื่อลดโอกาสการซ้ำซ้อนกับข้อมูลจริง

2. ควรจะเป็นตัวอักษรที่สามารถพิมพ์ได้ (printable character) เพื่อที่ในบางครั้งเราสามารถใช้โปรแกรมประเภท โปรแกรมบรรณาธิการ (text editor) แก้ไขและแทรกตัวอักขระแบ่งคำลงในข้อความได้โดยตรง

3. ตัวอักษรที่ใช้เป็นอักขระแบ่งคำ ควรจะมีลักษณะทางกายภาพที่สามารถแบ่งแยกจากตัวอักษรอื่นได้ง่ายและชัดเจน เพื่อให้สามารถเห็นขอบเขตของคำได้ชัดเจนขึ้น

จากคุณสมบัติข้างต้นจะเห็นว่าอักขระแบ่งคำชนิดตัวอักษรที่พิมพ์ไม่ได้ จะไม่เหมาะสมตามคุณสมบัติข้อ 2 และ 3 นอกจากนี้ตัวอักษรที่เป็นรหัสควบคุมอาจจะไปตรงกับรหัสควบคุมบางอย่าง ซึ่งอาจทำให้เกิดผลที่ไม่อาจคาดเดาได้

สำหรับอักขระแบ่งคำที่เป็นตัวอักษรแบ่งคำตามธรรมชาติ ก็มีคุณสมบัติเหมาะสมตามข้อ 2 และ 3 โดยเฉพาะลักษณะทางกายภาพของอักขระแบ่งคำจะเหมาะสมมาก เนื่องจากเป็นรูปแบบที่ผู้คนทั่วไปคุ้นเคยดีอยู่แล้ว อย่างไรก็ตาม อักขระแบ่งคำตามธรรมชาติก็ยังมีข้อเสียที่มักจะมีความถี่ของตัวอักษรสูง ทำให้มีโอกาสซ้ำกับข้อมูลจริงมาก

อักขระแบ่งคำชนิดที่ 3 คือตัวอักษรปกติ จะมีคุณสมบัติตามข้อ 2 คือเป็นตัวอักษรที่สามารถพิมพ์ได้ สำหรับคุณสมบัติข้อ 1 และ 3 นั้น เราสามารถเลือกตัวอักษรที่มีคุณสมบัติตรงตามนี้ได้

อันดับที่	ตัวอักษร	ความถี่	ความถี่สะสม	อันดับที่	ตัวอักษร	ความถี่	ความถี่สะสม
1	น	10.316	10.316	22	ผ	0.834	94.427
2	ร	8.961	19.277	23	ค	0.754	95.181
3	ก	7.096	26.373	24	ถ	0.749	95.930
4	อ	6.843	33.216	25	ณ	0.596	96.526
5	ง	6.607	39.823	26	ธ	0.559	97.085
6	ม	5.726	45.549	27	ญ	0.509	97.594
7	ย	4.829	50.378	28	ษ	0.468	98.062
8	ว	4.675	55.053	29	ภ	0.460	98.522
9	ล	3.906	58.959	30	ช	0.398	99.920
10	ท	3.873	62.832	31	ฐ	0.281	99.201
11	ด	3.853	66.685	32	ฟ	0.168	99.369
12	ต	3.387	70.072	33	ฉ	0.162	99.531
13	ห	3.368	73.440	34	ฝ	0.137	99.668
14	ส	3.114	76.554	35	ฏ	0.077	99.745
15	ป	3.100	79.654	36	ฎ	0.077	99.822
16	บ	2.865	82.519	37	ฒ	0.055	99.877
17	จ	2.771	85.290	38	ฮ	0.042	99.919
18	ค	2.481	87.771	39	ฆ	0.032	99.951
19	พ	2.061	89.832	40	ฑ	0.028	99.979
20	ข	2.013	91.845	41	ฬ	0.019	99.998
21	ช	1.748	93.593	42	ฌ	0.002	100.000

ตารางที่ 4.1 แสดงความถี่ของตัวอักษรไทย (เป็น เปอร์เซนต์)

จากตารางที่ 4.1 ซึ่งแสดงความถี่ของตัวอักษรในภาษาไทย (Tanaprakob 1984) จะเห็นว่าตัวอักษรที่มีความถี่ต่ำสุด 3 อันดับแรก คือ "ฌ", "ฬ" และ "ฑ" ตัวอักษรทั้ง 3 ตัวนี้จะมีคุณสมบัติเหมาะสมตามข้อ 1 ส่วนคุณสมบัติทางด้านกายภาพจะเห็นว่าตัว "ฬ" มีลักษณะเด่นต่างจากตัวอื่นที่เหลือ คือมีหางสูงขึ้นไปอีก 1 ระดับ ทำให้สามารถมองแยกออกจากข้อมูลอื่นได้ง่าย ทำให้ "ฬ" มีคุณสมบัติเหมาะสมที่จะใช้เป็นอักขระแบ่งคำ สำหรับใน

วิทยานิพนธ์ฉบับนี้จึงได้เสนอให้ใช้ "ฟ" เป็นอักขระแบ่งคำ

อย่างไรก็ตาม กฎเกณฑ์การเลือกอักขระแบ่งคำซึ่งกำหนดไว้นี้ ก็ไม่ได้เป็นกฎเกณฑ์ที่ตายตัว เราอาจจะเลือกอักษรที่ไม่มีคุณสมบัติตามนี้ แต่เหมาะสมกับงานของเราก็ได้ เช่น การใช้เว้นวรรคเป็นอักขระแบ่งคำ เพื่อให้ใช้งานกับโปรแกรมที่เขียนสำหรับข้อความภาษาอังกฤษได้ โดยยอมให้มีการซ้ำซ้อนของอักขระแบ่งคำกับข้อมูลจริงสูงขึ้น