บทที่ 2

พจนานุกรมอิเล็กทรอนิกส์

พจนานุกรมอิเล็กทรอนิกส์เป็นเสมือนแหล่งข้อมูลของงานด้านภาษาศาสตร์ที่จำเป็นสำหรับการประมวลผลภาษาธรรมชาติ [1] ภายในพจนานุกรมประกอบไปด้วยข้อสนเทศมากมายอาจเป็นข้อสนเทศย่อยๆ ซ้ำๆ กัน แต่ส่วนที่สำคัญของพจนานุกรมคือจะต้องเรียบเรียงข้อสนเทศในรูปแบบที่สะดวกในการค้นหาคำ [2] ส่วนรายละเอียดข้อสนเทศที่ปรากฏในพจนานุกรมอิเล็กทรอนิกส์จะแตกต่างกันไปตามแต่ประโยชน์การใช้งานของพจนานุกรมนั้นๆ เช่น พจนานุกรมอิเล็กทรอนิกส์ สำหรับงานด้านสัทศาสตร์จะประกอบด้วยรายละเอียดข้อมูลเสียงอ่านพร้อมกับชุดของคำศัพท์ที่อ่านออกเสียงตามคำอ่านดังกล่าว หรือพจนานุกรมอิเล็กทรอนิกส์สำหรับการประมวลผลภาษาธรรมชาติ รายละเอียดข้อมูลจะเป็นคำศัพท์ชนิดของคำศัพท์พร้อมความหมายของคำศัพท์ในแต่ละชนิด เป็นต้น

ประเภทของพจนานุกรมอิเล็กทรอนิกส์

ในปัจจุบันพจนานุกรมอิเล็กทรอนิกส์มีอยู่มากมาย แต่ไม่ได้จำแนกเป็นประเภทต่างๆ อย่างชัดเจน ผู้ใช้งานได้จัดหมวดหมู่เพื่อสร้างมาตรฐานสำหรับพจนานุกรมอิเล็กทรอนิกส์โดยจำแนกพจนานุกรมออกเป็นประเภทต่างๆ ตามหลักเกณฑ์ดังนี้

1. จำแนกตามประเภทผู้ใช้งาน

1.1 พจนานุกรมอเนกประสงค์ (General Dictionary) เช่น พจนานุกรมฉบับราชบัณฑิตยสถาน ปทานุกรมนักเรียน

1.2 พจนานุกรมศัพท์เทคนิค (Technical Dictionary) เก็บข้อมูลคำศัพท์เฉพาะสาขา เช่น สาขาการแพทย์, วิศวกรรมไฟฟ้า, ศัพท์คอมพิวเตอร์

1.3 พจนานุกรมเล็กซิคอน (User Specific Lexicon) เก็บข้อมูลคำศัพท์เฉพาะงานใดงานหนึ่งโดยเฉพาะเช่น พจนานุกรมสำหรับงานแปลภาษาด้วยคอมพิวเตอร์

2. จำแนกตามคู่ภาษา (Classify by Language Paires)

2.1 พจนานุกรมที่มีคู่ภาษาเพียงภาษาเดียว (Mono-Lingual Dictionary) เช่น พจนานุกรมฉบับราชบัณฑิตยสถาน พจนานุกรมของลองแมน (Longman Dictionary)

2.2 พจนานุกรมที่มีคู่ภาษา 2 ภาษา (Bilingual Dictionary) เช่น พจนานุกรมไทย-อังกฤษ ของ ส.เศรษฐบุตร

2.3 พจนานุกรมที่มีคู่ภาษามากกว่า 2 ภาษา (Multilingual Dictionary)

3. จำแนกตามเนื้อหาของข้อมูลที่บันทึกพจนานุกรม

3.1 เล็กซิคอน (Lexicon)

3.2 พจนานุกรม (Dictionary)

3.3 พจนานุกรมคำพ้อง (Thesaurus)

3.4 สารานุกรม (Encyclopedia)

4. จำแนกตามกลุ่มผู้ใช้งาน

4.1 พจนานุกรมสำหรับมนุษย์ (Dictionary for Human)

4.2 พจนานุกรมสำหรับเครื่องใช้ (Dictionary for Computer) ส่วนใหญ่ใช้เฉพาะงาน เช่น ฐานข้อมูลพจนานุกรมคำพ้อง (Database Retrieve Thesaurus)

รายละเอียดข้อมูลในพจนานุกรมอิเล็กทรอนิกส์

ดังที่ได้กล่าวในตอนต้นแล้วว่าพจนานุกรมอิเล็กทรอนิกส์มีมากมายหลายประเภท ขึ้นกับวัตถุประสงค์ของผู้ใช้งานเพราะฉะนั้นเนื้อหารายละเอียดที่บันทึกในพจนานุกรมก็แตกต่างกันไปในแต่ละงาน เช่นพจนานุกรมไทย-อังกฤษ ประกอบด้วยรายละเอียดข้อมูลคำศัพท์ไทย คำศัพท์ภาษาอังกฤษที่ความหมายสอดคล้องกับคำศัพท์ภาษาไทยพร้อมกับชนิดของคำศัพท์ว่าเป็น คำนาม คำกริยา คำสรรพนาม ส่วนพจนานุกรมคำพ้องนอกจากจะประกอบ

ด้วยคำศัพท์พร้อมความหมายแล้วจะต้องมีชุดของคำศัพท์ที่มีความหมายเหมือนกันด้วย ดังนี้ เป็นต้น

ดังนั้นหากว่ามีพจนานุกรมอิเล็กทรอนิกส์ที่เก็บรายละเอียดข้อมูลทุกอย่างที่สามารถนำไปใช้ในงานได้หลายๆ ด้านก็จะทำให้พจนานุกรมอิเล็กทรอนิกส์ชุดนั้นมีคุณค่านานับประการ ซึ่งจะเป็นประโยชน์ต่อกลุ่มผู้ใช้งานหลายกลุ่ม อีกทั้งสะดวกต่องานการประมวลผลข้อมูลทางด้านภาษาศาสตร์ที่ต้องนำพจนานุกรมมาประกอบการทำงานด้วย และที่สำคัญก็คือเป็นการสร้างมาตรฐานให้กับพจนานุกรมอีกด้วย

จะเห็นได้ว่าพจนานุกรมอิเล็กทรอนิกส์เป็นแหล่งรวมข้อมูลขนาดใหญ่ ที่ประกอบด้วยรายละเอียดมากมายที่เป็นประโยชน์สำหรับงานหลายๆ ด้าน สำหรับรายละเอียดข้อมูลที่น่าจะบันทึกไว้ในฐานข้อมูลพจนานุกรมอิเล็กทรอนิกส์มีดังนี้ [1]

ก. คำหรือหน่วยคำ
ข. ข้อมูลระดับคำ และการแบ่งแยกคำและพยางค์
ค. ข้อมูลเกี่ยวกับการผันของคำ
ง. รายละเอียดเกี่ยวกับการผันของคำ
จ. การออกเสียง
ฉ. ชนิดของคำ (Part of Speech)
ช. ความถี่ที่ใช้และอัตราความสำคัญของคำ
ซ. ข้อมูลเกี่ยวกับลักษณะของคำที่เป็นคำเก่าแก่ คำใหม่ คำที่ใช้ทั่วไป คำราชาศัพท์
ฌ. โครงสร้างทางไวยากรณ์ที่เกี่ยวข้อง
ญ. เฟรมของการก (Case of Frame)
ฎ. ตัวอย่างการใช้และประโยคตลอดจนรูปแบบประโยคที่ใช้
ฏ. คำพ้องความหมาย (Thesaurus)
ฐ. การเชื่อมโยงในคำ concept และไวยากรณ์

รายละเอียดเนื้อหาในพจนานุกรมอิเล็กทรอนิกส์ที่กล่าวมาในตอนต้นนั้น เป็นเพียงรายละเอียดที่ผู้วิจัยได้ทำการศึกษาจากเอกสารพบว่าจำเป็นสำหรับงานการประมวลผลข้อมูลด้านภาษาศาสตร์ แต่เนื่องจากผู้วิจัยมิได้ศึกษาวิชาภาษาศาสตร์โดยตรง อีกทั้งพิจารณาประโยชน์สำหรับการใช้งานบางอย่างเท่านั้น เพราะฉะนั้นอาจมีรายละเอียดบางอย่างที่ผู้อื่นเห็นสมควรว่าน่าจะจัดเก็บในพจนานุกรมอิเล็กทรอนิกส์ได้

โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์

โครงสร้างข้อมูลหมายถึงการจัดระเบียบและรูปแบบให้กับเขตข้อมูลต่างๆ ที่อยู่ในหน่วยความจำ [3] เพราะฉะนั้นในการประมวลผลข้อมูลด้วยคอมพิวเตอร์นั้นหากมีการประมวลผลกับข้อมูลที่ได้ออกแบบโครงสร้างที่เหมาะสมกับงานแล้วจะเอื้ออำนวยต่อการทำงานดังนี้

ก. สามารถนำข้อมูลเหล่านั้นออกมาใช้งานได้ทันทีตามความต้องการ

ข. ช่วยประหยัดเนื้อที่ของหน่วยความจำ เพื่อให้หน่วยความจำของคอมพิวเตอร์ถูกใช้งานอย่างมีประสิทธิภาพที่สุด

ค. สามารถค้นคืน (Retrieve) ข้อมูลได้อย่างรวดเร็ว

ง. สามารถปรับปรุงเปลี่ยนแปลงและพัฒนาโปรแกรมได้ง่าย

โครงสร้างข้อมูลที่รู้จักกันทั่วไปมีอยู่หลายประเภท โดยที่โครงสร้างข้อมูลแต่ละประเภทจะเหมาะสมสำหรับข้อมูลและการประมวลผลแบบหนึ่ง เพราะฉะนั้นในการที่จะเก็บข้อมูลจำเป็นต้องศึกษาทั้งลักษณะข้อมูล วิธีการนำไปใช้งาน ซึ่งหลักเกณฑ์ที่นำมาใช้สำหรับการพิจารณาโครงสร้างข้อมูล ประกอบด้วย

ก. จำนวนคำศัพท์ที่จัดเก็บในพจนานุกรม

ข. ขั้นตอนวิธีการค้นหาคำศัพท์

ค. ขั้นตอนวิธีการเพิ่มเติมแก้ไขข้อมูล

ง. เมื่อมีจำนวนคำศัพท์เพิ่มมากขึ้นจะมีผลกระทบต่อโครงสร้างข้อมูลอย่างไร

จ. วัตถุประสงค์ในการนำพจนานุกรมไปใช้งาน

พจนานุกรมอิเล็กทรอนิกส์ที่ใช้งานโดยทั่วไปในปัจจุบัน ได้ออกแบบโครงสร้างข้อมูลหลายๆ แบบ สำหรับการวิจัยครั้งนี้ผู้วิจัยได้ศึกษาโครงสร้างข้อมูลบางโครงสร้างที่ใช้งานบ่อยๆ โดยที่โครงสร้างข้อมูลที่กล่าวถึงมีลักษณะเด่นที่แตกต่างกันไป โครงสร้างข้อมูลที่ศึกษามีดังนี้

ก. โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียล (Index Sequential)

ข. โครงสร้างข้อมูลแบบบี-ทรี (B-Tree)

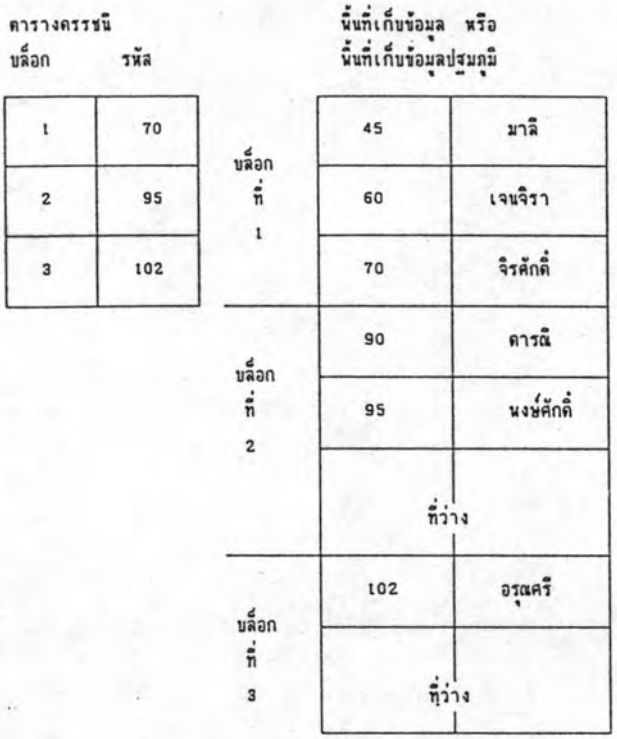
ค. โครงสร้างข้อมูลแบบทรัย (Trie)

ก. <u>โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียล</u>

เป็นโครงสร้างข้อมูลที่อาศัยวิธีการแบบอินเด็กซ์ชิง [4] [5] [6] [7] คือจะมีตารางดรรชนีสำหรับจัดเก็บรายละเอียดที่บอกให้ทราบว่าระเบียนที่ต้องการค้นหานั้นอยู่ส่วนใดของโครงสร้าง ตัวอย่างเช่น สมมุติว่าต้องการจัดเก็บข้อมูลในแฟ้มข้อมูลจำนวน n ระเบียน โดยที่ข้อมูลจัดเรียงลำดับตามคีย์ๆ หนึ่ง แต่ในหน่วยความจำหลักมีที่ว่างสำหรับเก็บข้อมูลเพียง x ระเบียน (x น้อยกว่า n) เพราะฉะนั้นถ้าต้องการจัดเก็บข้อมูลชุดนี้โดยใช้โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียลแล้วขั้นแรกจะต้องแยกข้อมูลชุดดังกล่าวเป็น n/k ส่วน ต่อจากนั้นจึงสร้างตารางดรรชนีสำหรับจัดเก็บคีย์ๆ พร้อมกับตำแหน่งของบล็อกที่เก็บข้อมูลชุดที่มีคีย์เป็นคีย์ดังกล่าว เพื่อนำไปใช้สำหรับการค้นหาข้อมูล รูปที่ 2.1 เป็นรูปแสดงลักษณะโครงสร้างข้อมูลแบบอินเด็กซีเควนเชียล

การสืบค้นข้อมูลในโครงสร้างข้อมูลแบบซีเควนเชียลนี้ เริ่มต้นทำการสืบค้นที่ตารางดรรชนี โดยที่นำคำที่ต้องการสืบค้นไปเปรียบเทียบกับคีย์ในตารางดังกล่าวว่าควรจะอยู่ที่บล็อกใด เมื่อทราบว่าอยู่บล็อกใดแล้วจึงไปดึงข้อมูลตามตำแหน่งที่ได้ ซึ่งขณะนี้จะได้ชุดของข้อมูลที่มีคีย์ใกล้เคียงกับคำที่ต้องการค้นหา ต่อจากนั้นจึงทำการสืบค้นคำที่ต้องการต่อไป

ตารางดรรชนีนั้นจะประกอบด้วย 2 ส่วน ส่วนแรกเป็นคีย์ที่ใช้สำหรับการทำการเปรียบเทียบ ซึ่งคีย์ที่เก็บอาจเป็นคีย์ที่มีค่าสูงสุดหรือต่ำสุดก็ได้ ส่วนที่ 2 เป็นตัวชี้ที่ชี้ตำแหน่งของบล็อก สำหรับตัวอย่างที่แสดงนี้เป็นแบบอินเด็กซ์ซีเควนเชียลระดับที่ 1 กรณีที่ข้อมูลมีจำนวนมาก จะทำให้ขนาดของตารางดรรชนีและบล็อกมีขนาดใหญ่ขึ้น ซึ่งสามารถกระทำโดยใช้อินเด็กซ์ซีเควนเชียลระดับที่ 2 หรือ อินเด็กซ์ซีเควนเชียลระดับที่ 3 ก็ได้ เพื่อใช้สำหรับเก็บตัวชี้ที่ชี้ไปยังตารางดรรชนีระดับที่ต่ำกว่าอีกที่หนึ่ง



รูปที่ 2.1 โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียล

เวลาที่ใช้สำหรับสืบค้น เป็นเวลาที่ใช้ในการสืบค้นในตารางดรรชนีรวมกับเวลาที่ใช้สืบค้นคำที่ต้องการจากบล็อก ซึ่งเวลาที่ใช้สำหรับสืบค้นในตารางดรรชนีจะน้อยมาก เนื่องจากตารางดังกล่าวเก็บในหน่วยความจำหลักและยังใช้วิธีการสืบค้นข้อมูลแบบทวิ แต่การสืบค้นในส่วนที่เก็บข้อมูลนั้นใช้การสืบค้นแบบลำดับ

โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียลนี้มีข้อด้อยตรงในกรณีที่มีการลบหรือเพิ่มข้อมูลหากการลบไม่ค่อยมีปัญหามากนักเพราะเราเพียงกำหนดแฟลกว่าระเบียนนั้นไม่ใช้ก็เพียงพอแล้ว แต่สำหรับการเพิ่มข้อมูลจะมีผลต่อขนาดของแฟ้มข้อมูลจนที่เก็บอาจไม่พอ หรือเพิ่มจนเนื้อที่ที่จัดเตรียมไว้ไม่พอต้องนำไปเก็บในเนื้อที่ส่วนขยาย (overflow area) ทำให้การปฏิบัติการช้าลง เพราะขนาดของข้อมูลใหญ่และข้อมูลส่วนใหญ่จัดเก็บในเนื้อที่ส่วนขยายนั่นเอง

<u>ข้อดีข้อเสียของโครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียล</u>

โครงสร้างข้อมูลแบบซีเควนเชียลเป็นโครงสร้างข้อมูลที่มีรูปแบบของโครงสร้างแบบง่ายๆ ไม่ซับซ้อน ค่าใช้จ่ายต่ำ ใช้สำหรับสื่อบันทึกข้อมูลหลายประเภท หากการเข้า

ถึงข้อมูลช้าจึงนิยมใช้โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียลเพื่อจัดเก็บข้อมูล ซึ่งสามารถเข้าถึงข้อมูลอย่างรวดเร็ว แต่ต้องใช้กับสื่อบันทึกข้อมูลที่มีการเข้าถึงโดยตรงเท่านั้น แต่ในการเก็บข้อมูลต้องคำนึงถึงบล็อกกิงแฟกเตอร์ด้วยก็คือ จำนวนระเบียนที่จัดเก็บต่อหนึ่งบล็อก ซึ่งมีความสำคัญต่อความเร็วในการเข้าถึงข้อมูล ถ้าสามารถจัดให้ขนาดของบล็อกมีขนาดใกล้เคียงกับขนาดของแทร็คแล้วจะทำให้เวลาที่ใช้สำหรับการเข้าถึงยิ่งสั้น นอกจากนี้ต้องคำนึงถึงระดับของดรรชนีอีกด้วย หากมีดรรชนีหลายระดับจะทำให้สิ้นเปลืองเวลาในการสืบค้นในส่วนของตารางดรรชนีมากขึ้น แต่ถ้าระดับน้อยไปก็จะทำให้เสียเวลาสืบค้นในส่วนที่เก็บข้อมูลมากเช่นกัน นอกจากนี้เมื่อมีการเพิ่มหรือลบข้อมูลมาก ๆ แล้วต้องดำเนินการปรับโครงสร้างใหม่เพื่อนำเนื้อที่ที่ไม่ได้ใช้มาใช้งานใหม่

ข. <u>โครงสร้างข้อมูลแบบบี-ทรี</u>

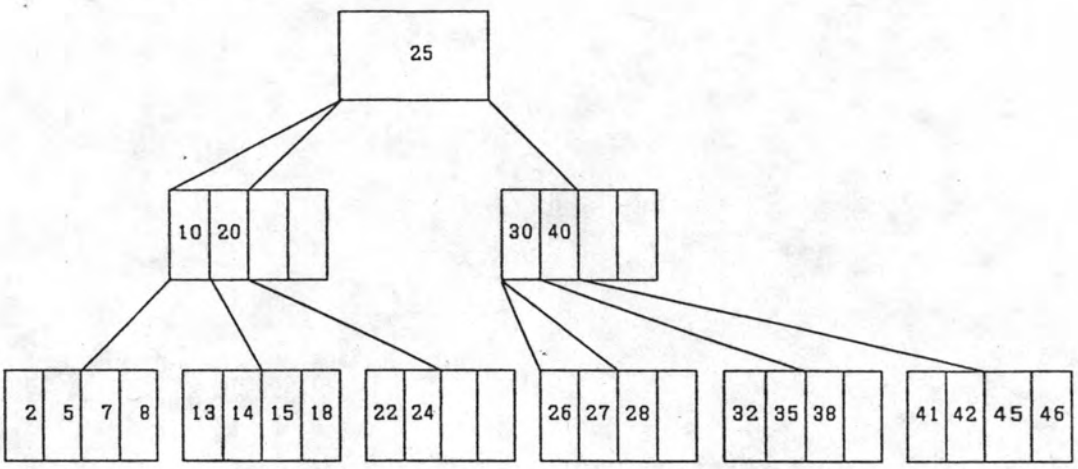
โครงสร้างข้อมูลแบบบี-ทรีนั้นเป็นโครงสร้างข้อมูลแบบทรีประเภทหนึ่ง [4] [5] [6] [7] ซึ่งมีลักษณะพิเศษดังนี้ สำหรับลักษณะโครงสร้างข้อมูลแบบบี-ทรีอันดับ n ก็ต่อเมื่อ

1. โหนดแต่ละโหนดของบี-ทรี ประกอบด้วยคีย์ทั้งหมดจำนวน k คีย์ โดยที่ $n \leq k \leq 2n$ และคีย์ที่จัดเก็บในโหนดนั้นเรียงลำดับน้อยไปมากจากทางซ้ายไปทางขวาของโหนด

2. ภายในโหนดจะมีตัวชี้ (pointer) ที่ชี้ไปยังโหนดลูก (son) อีกจำนวน n+1 ตัวชี้ หากโหนดใดที่ตัวชี้ทุกๆ ตัวชี้ไปที่นัล (NULL) แล้ว เราจะเรียกโหนดดังกล่าวว่าโหนดสิ้นสุด (Terminate Node)

3. โหนดสิ้นสุดของโครงสร้างข้อมูลแบบบี-ทรีนี้จะอยู่ที่ระดับ (Level) เดียวกันทั้งสิ้น

รูปที่ 2.2 เป็นตัวอย่างโครงสร้างข้อมูลแบบบี-ทรีอันดับ 2 ที่มี 3 ระดับ



รูปที่ 2.2 โครงสร้างข้อมูลแบบบีทรี

<u>การเพิ่มข้อมูลในโครงสร้างข้อมูลแบบบี-ทรี</u>

1. รูปที่ 2.3 แสดงลักษณะโหนดที่มีจำนวนคีย์เท่ากับ m คีย์ของบี-ทรี

$$p_0 k_1 p_1 k_2 p_2 \ \cdots \ p_{m-1} k_m p_m$$

p = ตัวชี้ที่ชี้ไปยังโหนดลูก และ k คือคีย์

รูปที่ 2.3 ลักษณะโหนดของโครงสร้างข้อมูลแบบบี-ทรี

2. ถ้าต้องการเพิ่ม x ลงในบี-ทรี
3. สืบค้นในแต่ละโหนดโดยเริ่มต้นที่โหนดราก (root node) ซึ่งมีขั้นตอนการสืบค้นดังนี้

   3.1 ถ้า $k_i < x < k_{i+1}$; $1 \leq i \leq m$ ต้องไปค้นหาต่อในโหนดที่ชี้โดย ตัวชี้ p_i

3.2 ถ้า $k_m < x$ ต้องไปค้นหาต่อในโหนดที่ชี้โดย ตัวชี้ p_m

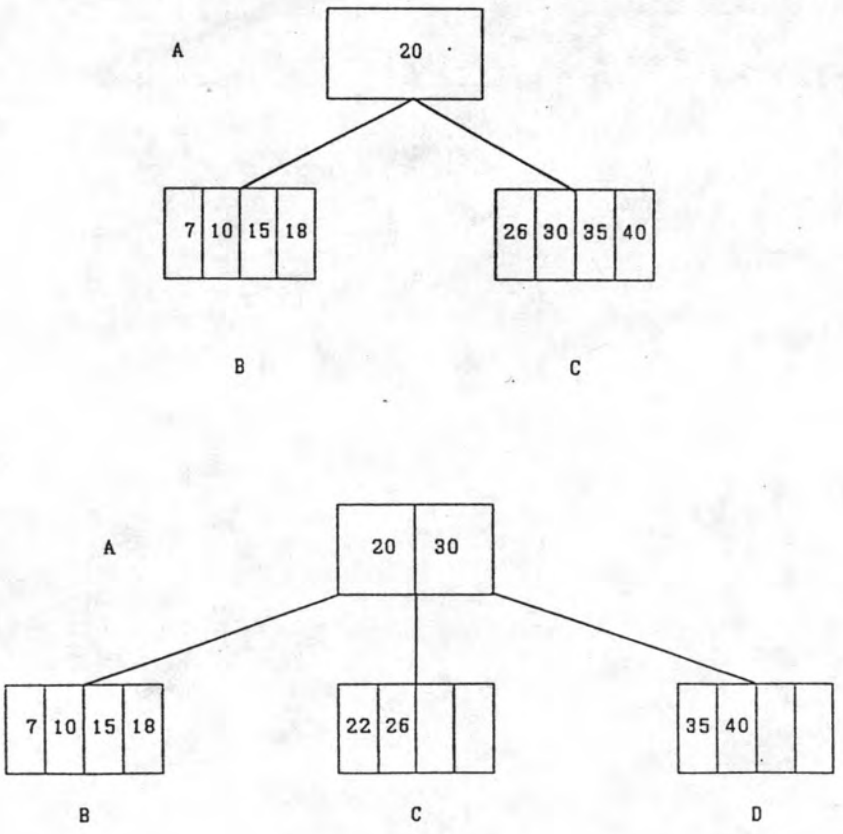
3.3 ถ้า $x < k_1$ ต้องไปค้นหาต่อในโหนดที่ชี้โดย ตัวชี้ p_0

4. กรณีที่ทำการสืบค้นจนถึงโหนดสิ้นสุดแล้วให้เพิ่ม x ในโหนดดังกล่าว ดังนี้

4.1 ถ้าโหนดสิ้นสุดมีเนื้อที่ว่างสำหรับใส่ค่า x ให้ใส่คีย์ดังกล่าว โดยจัดเรียงจากน้อยไปมาก

4.2 ถ้าโหนดสิ้นสุดเต็มให้ใช้เทคนิคที่เรียกว่าการแยกโหนด โดยย้ายคีย์ที่เก็บไว้ในตำแหน่งที่ m/2 ไปเก็บไว้ในโหนดที่อยู่ในระดับสูงกว่า (โหนดก่อนหน้า) พร้อมทั้งแยกโหนดดังกล่าวออกเป็นโหนดใหม่ 2 โหนดโดยที่นำสมาชิกตัวที่ 1 ถึง m/2 -1 เก็บในโหนดทางซ้ายมือ และเก็บสมาชิกตัวที่ m/2 + 1 ถึง ตัวที่ m เก็บในโหนดทางขวามือ โดยจัดเรียงจากน้อยไปมาก

<u>ตัวอย่างการเพิ่มข้อมูลในบี-ทรี</u>



รูปที่ 2.4 ตัวอย่างการเพิ่มเลข 22 ในบี-ทรี

สมมุติว่าต้องการเพิ่มค่า 22 ในบี-ทรี พบว่าต้องใส่ค่า 22 ในโหนด C แต่โหนด C เต็มแล้วเพราะฉะนั้นต้องแบ่งโหนด C ออกเป็น 2 โหนดในที่นี้คือโหนด C และโหนด D พร้อมทั้งนำคีย์ตัวที่ m/2 คือ 30 ไปใส่ในโหนด A ซึ่งอยู่ในระดับที่สูงกว่า C ดังรูป 2.4 (ข)

ข้อดีข้อเสียของบี-ทรี

ข้อดีของบี-ทรีคือเป็นโครงสร้างข้อมูลที่ใช้เก็บข้อมูลในสื่อบันทึกข้อมูลประเภทดิสก์ ซึ่งเวลาที่ใช้ในการสืบค้นไม่มากนัก อีกทั้งมีการแยกโหนดในขณะที่ทำการเพิ่มคีย์โดยอัตโนมัติ ตลอดจนการเก็บข้อมูลไม่จำเป็นต้องรู้ถึงลักษณะทางกายภาพของสื่อบันทึกข้อมูล

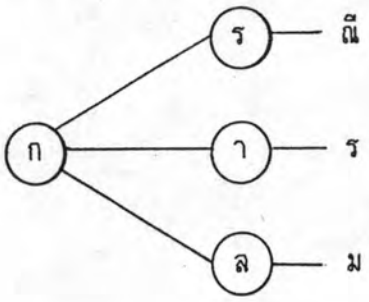
ส่วนข้อเสียของบี-ทรี ก็คือข้อมูลมีการเคลื่อนย้ายบ่อยครั้งอันเนื่องมาจากการแยกโหนดในขณะที่มีการเพิ่มหรือลบคีย์นั่นเอง อีกประการหนึ่งก็คือการกำหนดขนาดของโหนดจะต้องนำจำนวนระเบียนที่จัดเก็บในโหนดมาเป็นส่วนร่วมในการกำหนดด้วย

ค. โครงสร้างข้อมูลแบบทรัย

โครงสร้างข้อมูลแบบทรัยเป็นโครงสร้างข้อมูลที่มีประกอบด้วยโหนดต่างๆ [9] [10] [11] เช่นเดียวกันกับโครงสร้างข้อมูลแบบทรี แต่วิธีการเก็บข้อมูลในโครงสร้างข้อมูลจะแตกต่างกันคือ โครงสร้างข้อมูลแบบทรัยใช้ตัวอักษรของคำศัพท์ในการดำเนินการ ส่วนโครงสร้างข้อมูลแบบทรีใช้คำศัพท์ทั้งคำในการเปรียบเทียบและสร้างโหนดต่างๆ

โครงสร้างข้อมูลแบบทรัยประกอบด้วยโหนดต่างๆ ซึ่งสร้างจากตัวอักษรกับตัวชี้ที่ชี้ไปยังโหนดที่เป็นโหนดลูก ดังแสดงตามรูป 2.5 โครงสร้างข้อมูลแบบทรัยนี้เป็นโครงสร้างข้อมูลที่เหมาะสำหรับจัดเก็บพจนานุกรม เนื่องจากเป็นโครงสร้างข้อมูลที่มีประสิทธิภาพทางด้านพรีฟิกซ์เสิชชิง (Prefix Searching) คือเป็นวิธีการสืบค้นในลักษณะที่ผลลัพธ์ของการสืบค้นเป็นชุดคำศัพท์ที่มีส่วนของตัวอักษรที่ขึ้นต้นเหมือนกันเช่น การสืบค้นชุดของคำศัพท์ comput* โดยที่เครื่องหมาย '*' ใช้แทนว่าส่วนที่เหลือของคำศัพท์เป็นอะไรก็ได้ เพราะฉะนั้นหลังจากการสืบค้นจะใช้คำศัพท์ compute computation computer เป็นต้น แต่อย่างไรก็ตามปัญหาที่เกิดขึ้นในการสืบค้นคำศัพท์โดยวิธีนี้ก็คือ กรณีที่จำนวนตัวอักษรที่ขึ้นต้นเหมือนกันมากแล้วจำนวนโหนดที่ต้องสืบค้นจะมีจำนวนมากด้วย จะมีผลทำให้เวลาที่พบโหนดที่แยกความแตกต่างของคำศัพท์นานขึ้น

สำหรับเวลาที่ใช้ในการสืบค้นเฉลี่ยของโครงสร้างข้อมูลแบบทรัยของคำศัพท์ m คำที่ประกอบด้วยโหนด n โหนดเท่ากับ $\log_m N$



ⓚ - กระดาษ
ⓑ - ขจัด
ⓟ - ประเพณี

รูปที่ 2.5 โครงสร้างข้อมูลแบบทรัย

จากโครงสร้างข้อมูลทั้ง 3 รูปแบบนี้ เป็นโครงสร้างข้อมูลที่มีข้อเด่นและข้อด้อยต่างๆ กัน เช่นโครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียลนั้นจะใช้เนื้อที่ในการจัดเก็บข้อมูลไม่มากนัก อีกทั้งการบันทึกข้อมูลตลอดจนการค้นหาข้อมูลไม่ซับซ้อน ในขณะที่โครงสร้างข้อมูลแบบบี-ทรีโครงสร้างค่อนข้างซับซ้อนวิธีการบันทึกข้อมูลและวิธีการค้นหาข้อมูลยุ่งยากยิ่งขึ้น แต่สามารถค้นหาข้อมูลได้รวดเร็วกว่าการค้นหาข้อมูลในโครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียล ในกรณีที่จำนวนคำศัทพ์ที่บันทึกในพจนานุกรมอิเล็กทรอนิกส์มีปริมาณมากๆ จะใช้เวลาในการค้นหามากขึ้นสำหรับโครงสร้างข้อมูลแบบทรัยนั้นนับว่าสิ้นเปลืองเนื้อที่ค่อนข้างมากแต่วิธีการบันทึกตลอดจนการค้นหาข้อมูลยุ่งยากยิ่งขึ้น แต่ความเร็วในการค้นหาข้อมูลจะขึ้นอยู่กับความยาวของคำศัพท์มิใช่จำนวนคำศัพท์ที่บันทึกในพจนานุกรมเหมือนกับการค้นหาคำศัพท์ในโครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเชียลและบี-ทรี

ผลงานการสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่สนับสนุนการประมวลผลภาษาไทยด้วยเครื่องคอมพิวเตอร์

ปัจจุบันมีนักวิจัยหลายท่านที่ให้ความสนใจมุ่งพัฒนาโครงสร้างสำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย เพื่อสนับสนุนงานการประมวลผลภาษาไทยด้วยคอมพิวเตอร์อันได้แก่

Data Compression Techniques for ElecTronic Dictionary

งานวิจัยชิ้นนี้เป็นงานวิจัยของ รศ.ยืน ภู่วรวรรณ และ ผศ.ดร.ชัยยงค์ วงศ์สุรวัฒน์ [12] โดยตีพิมพ์ในเอกสารประกอบการสัมมนาทางวิชาการ Proceeding of The Regional Workshop on Computer Processing of Asian Language (CPAL) เป็นงานวิจัยเพื่อออกแบบและพัฒนาพจนานุกรมอิเล็กทรอนิกส์ (Electronic Dictionary) พจนานุกรมดังกล่าวประกอบด้วยคำศัพท์ภาษาไทยประเภทของคำศัพท์ (category) และความหมายของคำศัพท์พจนานุกรมดังกล่าวคำนึงถึงวิธีการเข้าถึงที่รวดเร็วและสามารถใช้งานได้อย่างมีประสิทธิภาพ โดยที่ผู้วิจัยได้เสนอรูปแบบโครงสร้างของลักษณะข้อมูลสำหรับการจัดทำพจนานุกรมตรวจสอบตัวสะกดนี้ 3 รูปแบบ ทั้งนี้พจนานุกรมนอกจากจะใช้ตรวจสอบตัวสะกดแล้วยังสามารถเพิ่มเติมรายละเอียดอื่นๆ ที่ต้องการ เพื่อนำไปใช้ในงานการวิเคราะห์ข้อความทางภาษาด้วย

1. โครงสร้างพจนานุกรม

พจนานุกรมที่ทำการพัฒนาประกอบด้วย 2 ส่วน คือ ส่วนที่ใช้เก็บคำศัพท์เพื่อใช้ในการตรวจสอบตัวสะกด และส่วนที่ใช้เก็บสารสนเทศทางภาษาศาสตร์ เช่น ประเภทของคำศัพท์ (category) ความหมายเป็นภาษาอังกฤษ เป็นต้น

โครงสร้างที่นักวิจัยทำการออกแบบสำหรับเก็บคำศัพท์เพื่อใช้ในการตรวจสอบตัวสะกดมี 3 รูปแบบคือ

ก. โครงสร้างแบบเทเบิลเทเบิลลิเนียเสิช (Table-Table-Linear-Search (TTLS))

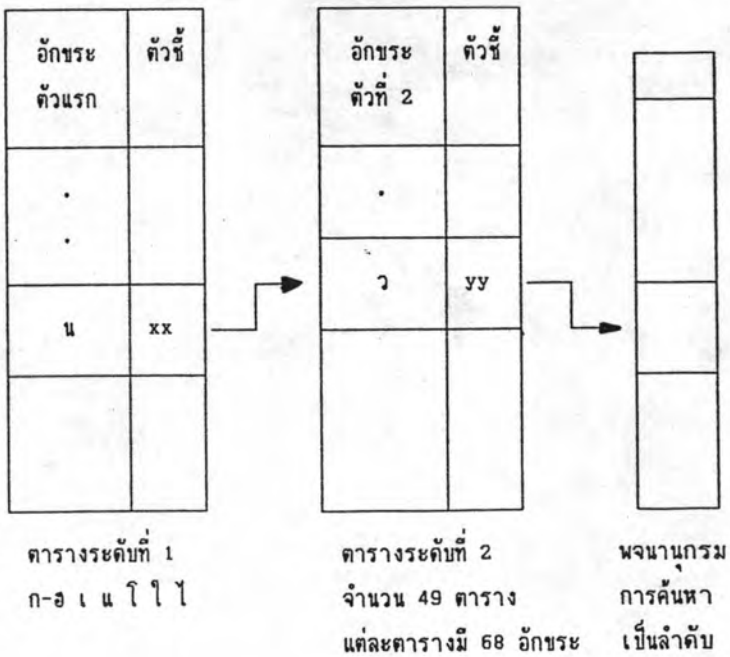
การเก็บคำศัพท์ในพจนานุกรมถ้าต้องการประหยัดเนื้อที่การเก็บจะออกแบบการเก็บในลักษณะที่คำศัพท์วางเรียงติดต่อกันไป โดยมีตัวเลขบอกจำนวนตัวอักขระที่ตามมาอยู่ข้างหน้าชุดของอักขระดังนี้

ก....2กา3กาก3กาง6กางเกง3กาจ5กาชาด.....

การเก็บคำศัพท์ในลักษณะดังกล่าวพบว่ามีตัวอักขระบางตัวซ้ำกัน ซึ่งสิ้นเปลืองเนื้อที่โดยใช่เหตุ ดังนั้นจึงแยกตัวอักขระ 2 ตัวแรกมาเก็บในตารางต่างหาก แล้วมีตัวชี้ชี้ไปยังหมายเลขระเบียน (record) ในพจนานุกรม ซึ่งพจนานุกรมจะเก็บอักขระส่วนที่เหลือถัดจากอักขระ 2 ตัวแรก เพราะฉะนั้นในพจนานุกรมจะมีข้อมูลดังนี้

01ก1ง4งเกง1จ3ชาด.....

โดยที่รายละเอียดของข้อมูลที่จัดเก็บ คือ ตัวเลขที่บอกจำนวนอักขระและอักขระจำนวนเท่ากับตัวเลขที่ระบุ รูปที่ 2.6 แสดงโครงสร้างข้อมูลแบบทีทีเอส



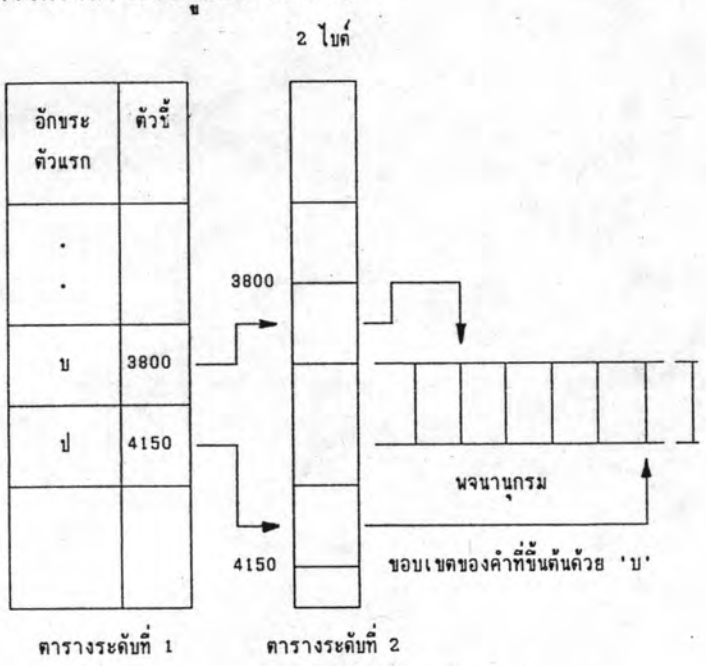
รูปที่ 2.6 แสดงโครงสร้างข้อมูลแบบทีทีเอส

ข. โครงสร้างแบบเทเบิลเทเบิลอินเด็กซ์เสิช (Table-Index-Search (TIS))

โครงสร้างข้อมูลแบบนี้ถูกออกแบบโดยใช้หลักการที่จะทำให้ขนาดการใช้เนื้อที่ในพจนานุกรมจะน้อยลง โดยเก็บคำศัพท์ถูกเก็บไว้ ดังนี้

....2กา3กาก3กาง6กางเกง3กาจ5กาชาด.....

โครงสร้างข้อมูลแบบนี้ประกอบด้วยตารางสืบค้นมีลักษณะคล้ายตารางสืบค้นของโครงสร้างแบบที่ 1 แต่ตารางระดับที่ 1 เก็บอักขระตัวแรกของคำศัพท์ ส่วนตารางระดับที่ 2 เป็นตารางที่เก็บตัวชี้ที่ชี้ไปยังตำแน่งที่เก็บคำศัพท์ในพจนานุกรม รูปที่ 2.7 แสดงโครงสร้างข้อมูลแบบทีไอเอส

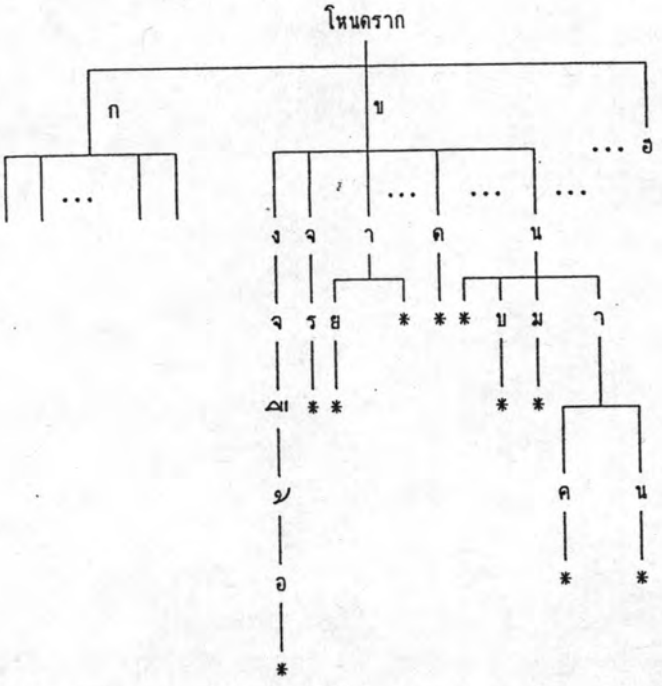


รูปที่ 2.7 แสดงโครงสร้างข้อมูลแบบทีไอเอส

ค. โครงสร้างข้อมูลแบบต้นไม้ (Tree Structure)

โครงสร้างข้อมูลแบบนี้ประกอบด้วยโหนด (node) ต่างๆ และโหนดเหล่านั้นเป็นที่เก็บตัวอักขระ พร้อมทั้งมีตัวชี้ชี้ไปยังโหนดอื่นดังแสดงใน รูปที่ 2.8

วิธีการเก็บข้อมูลในแฟ้มข้อมูล ข้อมูลจะถูกแยกเก็บเป็นกลุ่มๆ โดยใช้อักขระตัวแรกของคำศัพท์ในการจัดกลุ่ม



รูปที่ 2.8 โครงสร้างข้อมูลแบบต้นไม้

2. <u>โครงสร้างข้อมูลสำหรับเก็บประเภทและความหมายของคำศัพท์</u>

โครงสร้างส่วนที่ 2 ของพจนานุกรม ใช้โครงสร้างแบบทรีเนื่องจากคำศัพท์ภาษาไทยเป็นคำประสม เช่น คำว่า 'ช่อง' ปกติแล้วคำนี้ก็มีความหมายในตัวเองอยู่แล้วแต่ถ้านำไปประสมกับคำอื่นก็จะได้ความหมายใหม่ ได้แก่ 'ช่องแคบ' ,'ช่องเขา', 'ช่องไฟ' ดังนั้นจึงสร้างเป็นทรี 3 ระดับ ระดับแรกเก็บคำโดด ในที่นี้คือ'ช่อง' ระดับที่ 2 จะมีกลุ่มของคำประสมที่เกิดจากการประสมของ 'ช่อง' ส่วนในระดับที่ 3 เป็นระดับที่เก็บประเภท (category) และความหมายภาษาอังกฤษของคำประสมเหล่านั้น หากคำประสมนั้นมีหลายความหมายก็จัดเก็บความหมายเหล่านั้นไว้ทั้งหมด

3. <u>สรุปผลงานวิจัย</u>

ในการวิจัยนี้ผู้วิจัยได้สร้างพจนานุกรมที่บรรจุคำศัพท์ 18,569 คำ และวิสามานยนาม (proper noun) อีก 5000 คำ ตารางที่ 2.1 เป็นตารางแสดงประสิทธิภาพการเก็บคำศัพท์ในโครงสร้างต่าง ๆ จากการวิจัยนี้

	จำนวนเนื้อที่ที่ใช้ (ไบท์)	ความเร็วเฉลี่ยที่ใช้ในการตรวจสอบตัวสะกดของเอกสารขนาด 29 บรรทัด (วินาที)
ทีทีแอลเอส	124,608	48
ทีไอเอส	188,888	23
แบบต้นไม้	94,549	41

ตารางที่ 2.1 ตารางแสดงการเปรียบเทียบประสิทธิภาพของโครงสร้างข้อมูลแบบต่าง ๆ (พจนานุกรมบรรจุคำศัพท์ 18,569 คำ)

ทีทีแอลเอส = โครงสร้างข้อมูลแบบเทเบิลเทเบิลลินเนียเสิช
ทีไอเอส = โครงสร้างข้อมูลแบบเทเบิลอินเด็กซเสิช

จากงานวิจัยที่ได้กล่าวมาแล้วข้างต้นนั้น โครงสร้างที่ใช้ในการจัดเก็บพจนานุกรมแต่ละแบบยังมิได้นำเอาลักษณะของคำภาษาไทยมาเป็นส่วนหนึ่งของหลักเกณฑ์การพิจารณาโครงสร้างข้อมูล แต่โครงสร้างข้อมูลแบบทรัยที่สนองหลักการดังกล่าว อีกทั้งหากจำนวนคำศัพท์ที่จัดเก็บในฐานข้อมูลพจนานุกรมอิเล็กทรอนิกส์มีจำนวนมาก ไม่สามารถจัดเก็บในหน่วยความจำหลักได้เพียงพอแล้วจะหาโครงสร้างใดที่รองรับหลักเกณฑ์ดังกล่าวได้ และนอกจากนี้เนื้อที่ที่ใช้สำหรับการจัดเก็บคำศัพท์จะต้องไม่สูญเสียในการเก็บข้อมูลว่างๆ ซึ่งน่าจะมีวิธีที่ใช้เนื้อที่สำหรับเก็บคำศัพท์อย่างมีประสิทธิภาพสูงสุด ซึ่งในบทต่อไปจะกล่าวถึงโครงสร้างข้อมูลที่ผู้วิจัยได้ทำการศึกษาและพิจารณาว่าเหมาะสมที่จะใช้สำหรับค้นหาคำศัพท์ที่ไม่ทราบถึงตำแหน่งสิ้นสุดของคำศัพท์ที่แน่นอนดังเช่นคำศัพท์ภาษาไทยได้อย่างรวดเร็ว อีกทั้งใช้เนื้อที่ในการเก็บคำศัพท์อย่างมีประสิทธิภาพด้วย

สำหรับโครงสร้างข้อมูลพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่นำเสนอในวิทยานิพนธ์นี้จะทำการทดสอบเปรียบเทียบกับโครงสร้างข้อมูลจากงานวิจัยของ ร.ศ. ยืน ภู่วรวรรณ และจะนำเสนอรายละเอียดการทดสอบ และสรุปผลการทดสอบไว้ในบทที่ 6 และ บทที่ 7