



ปริทัศน์วรรณกรรม

การปริทัศน์วรรณกรรมในบทนี้ มีเนื้อหาเกี่ยวกับวรรณคดีในหัวข้อต่าง ๆ คือ 1) คำจำกัดความของวรรณคดี 2) ขั้นตอนการหาวรรณคดี 3) ภาษาวรรณคดี 4) การหาวรรณคดีระบบอัตโนมัติ 5) การหาวรรณคดีภาษาไทยระบบอัตโนมัติ 6) การวัดประสิทธิภาพของวรรณคดี 7) การเปรียบเทียบคุณลักษณะของศัพท์ควบคุมและศัพท์อิสระ และ 8) การค้นคืนด้วยศัพท์ควบคุมและศัพท์อิสระ ทั้งนี้ได้แทรกงานวิจัยไว้ในหัวข้อต่าง ๆ ที่เกี่ยวข้อง

วรรณคดีนั้นนับได้ว่ามีบทบาทสำคัญในการจัดเก็บและการค้นคืนสารสนเทศ การดำเนินการจัดเก็บสารสนเทศ ประกอบด้วยกิจกรรมหลายประการ กิจกรรมสำคัญประการหนึ่งคือ การกำหนดคำศัพท์แทนเนื้อหาของเอกสารหรือเรียกว่า การหาวรรณคดี วรรณคดีที่จัดทำขึ้นนี้มีบทบาทในการค้นคืนสารสนเทศ เนื่องจากเป็นเครื่องมือทางบรรณานุกรมที่ช่วยชี้ตำแหน่งของข้อมูลที่อยู่ในสิ่งพิมพ์หรือชี้ตำแหน่งของสิ่งพิมพ์ที่มีอยู่ในห้องสมุดหรือศูนย์เอกสาร (ชลัษเฐียร เหมะรัชตะ, 2535: 20) ดังนั้น การจัดทำวรรณคดีจึงมีความสัมพันธ์และเกี่ยวโยงกับการค้นหาสารสนเทศ กล่าวคือประสิทธิภาพของการค้นคืนข้อมูล ส่วนหนึ่งจะขึ้นอยู่กับคุณภาพของวรรณคดี

คำจำกัดความของวรรณคดี

ตามคำจำกัดความของสถาบันมาตรฐานแห่งชาติอเมริกัน (American National Standards Institute) ตามมาตรฐาน ANSI 1968 วรรณคดี คือ เครื่องมือที่จัดทำขึ้นอย่างมีระบบ ที่นำไปสู่รายการหรือแนวความคิดของกลุ่มเอกสาร รายการและแนวความคิดดังกล่าวได้รับการแทนที่ด้วยรายการคำสั่งจัดเรียงตามลำดับที่สามารถค้นได้ เช่น เรียงตามลำดับอักษรตามลำดับเหตุการณ์ หรือหมายเลข (Borko and Bernier, 1978: 8)

ในหนังสือ The ALA Glossary of Library and Information Science วรรณคดี หมายถึง เครื่องมือที่จัดทำขึ้นอย่างมีระบบซึ่งนำไปสู่เนื้อหาของเอกสาร ประกอบด้วย คำศัพท์ รายการอ้างอิง หมายเลขเอกสาร รายการเลขหน้า และอื่น ๆ ที่มีการจัดเรียงตามลำดับ เพื่อให้ผู้ใช้เข้าถึงเนื้อหาของเอกสารที่ต้องการได้อย่างรวดเร็ว (Young, 1983: 116)

Rowley (1988: 48) ได้ให้คำจำกัดความของวรรณคดีว่า วรรณคดีคือชุดของหัวเรื่องหรือจุดเข้าถึง (Entry points) ซึ่งมีการจัดเรียงตามลำดับที่เป็นที่ยอมรับ จุดเข้าถึงแต่ละรายการจะแสดงว่าข้อมูลเพิ่มเติมหรือรายละเอียดของเอกสารนั้น ๆ สามารถหาได้จากที่ใด

จากคำจำกัดความข้างต้น สรุปได้ว่าดัชนี คือรายการศัพท์ที่จัดเรียงตามลำดับ เป็นคำที่ใส่แทนเนื้อหาของเอกสาร หน้าที่ชี้ตำแหน่งของเนื้อหาเอกสาร

ขั้นตอนการทาดัชนี

การทาดัชนี เป็นวิธีการหนึ่งของการแทนที่เอกสาร (Information representation) การจัดทาดัชนีในที่นี้คือการทาดัชนีเรื่อง (Subject indexing) ซึ่งในการจัดทาดัชนีประกอบด้วยขั้นตอนสำคัญ 2 ขั้นตอน คือ ขั้นตอนที่ 1 เป็นการวิเคราะห์แนวความคิดหรือเนื้อหาของเอกสาร ขั้นตอนที่ 2 เป็นการแปลผลการวิเคราะห์แนวความคิดเป็นคำศัพท์ (Lancaster, 1979: 9)

ในขั้นตอนการวิเคราะห์เนื้อหาหรือแนวความคิดของเอกสาร ผู้ทาดัชนีจะต้องเข้าใจเนื้อหาของเอกสารและรู้ความต้องการของผู้ใช้เป็นอย่างดี เพราะเป็นขั้นตอนที่จะต้องพิจารณา คัดเลือกประเด็นสำคัญในเอกสาร หลักเกณฑ์ในการคัดเลือกเรื่องมาทาดัชนี เป็นเรื่องยากที่จะกำหนดเป็นลายลักษณ์อักษร ต้องอาศัยประสบการณ์และวิจารณญาณของผู้ทาดัชนี (Rowley, 1988: 51)

ขั้นตอนที่ 2 เป็นการเลือกศัพท์ที่เหมาะสมมาแทนเรื่อง ที่ได้คัดเลือกมาทาดัชนี ซึ่งหมายถึงการแปลการวิเคราะห์แนวความคิดให้อยู่ในรูปคำศัพท์ หรือภาษาของดัชนี ระบบดัชนีส่วนใหญ่จะใช้ศัพท์ควบคุมซึ่งเป็นชุดของศัพท์จำนวนหนึ่งที่ใช้แทนเนื้อหาของเอกสาร ศัพท์ดังกล่าวอาจเป็นรายการหัวเรื่อง แผนการกำหนดหมวดหมู่ ศัพท์สัมพันธ์ หรือรายการคำหรือวลีที่ได้รับการยอมรับแล้ว ส่วนศัพท์ไม่ควบคุมจะไม่มีข้อกำหนดเกี่ยวกับศัพท์ที่ผู้ทาดัชนีเลือกใช้ ส่วนใหญ่จะใช้คำหรือวลีที่ปรากฏในเอกสาร ศัพท์ที่ผู้ทาดัชนีใช้แทนเนื้อหาเอกสาร ไม่ว่าจะเป็นศัพท์ควบคุม หรือไม่ควบคุม ในที่นี้จะเรียกว่า ศัพท์ดัชนี

ภาษาดัชนี

ภาษาดัชนี คือภาษาที่ใช้อธิบายเรื่องหรือคุณลักษณะอื่น ๆ ของข้อมูล หรือเอกสาร ที่นำมาจัดทาดัชนี (Rowley, 1988: 52) ภาษาดัชนีแบ่งได้ 2 ประเภท คือศัพท์ควบคุม หรือภาษาควบคุม (Controlled vocabulary หรือ Controlled language) และศัพท์อิสระ หรือภาษาธรรมชาติ (Uncontrolled vocabulary หรือ Natural language)

ศัพท์ควบคุม

ศัพท์ควบคุม (Controlled vocabulary) คือ ชุดของศัพท์จำนวนหนึ่งที่ได้รับบริการยอมรับ และมีการกำหนดขึ้นอย่างมีแบบแผน มีการจัดเรียงตามลำดับที่สามารถเข้าถึงได้ นอกจากนี้ อาจมีการแสดงความสัมพันธ์ระหว่างคำในรายการ เช่น มีรายละเอียดเกี่ยวกับการ

โยงคำ หรือข้อมูลที่แสดงถึงความสัมพันธ์ในลักษณะกว้างกว่าหรือแคบกว่าของคำแต่ละคำ เป็นต้น ผู้ทําการจะพิจารณาคัดเลือกคำศัพท์ที่เหมาะสมจากรายการศัพท์ควบคุมเพื่ออธิบายหรือแทนที่เนื้อหาของเอกสาร ในทำนองเดียวกัน ผู้ค้นข้อมูลจะต้องกำหนดคำศัพท์แทนเรื่องที่ต้องการค้นหาให้ตรงกับคำศัพท์ในรายการศัพท์ควบคุม ตัวอย่างศัพท์ควบคุม ได้แก่ แผนการกำหนดหมวดหมู่เพื่อการจัดเก็บหนังสือชั้น (Lancaster, 1989: 2) รายการหัวเรื่อง เช่น Sears' List of Subject Headings, Library of Congress Subject Headings, Subject Headings in Engineering และคำศัพท์ในสาขาวิชาต่าง ๆ (Rowley, 1988: 52)

การใช้ศัพท์ควบคุมในการทําการค้น มีวัตถุประสงค์สำคัญ 2 ประการ คือ

1. เพื่อลดความซ้ำซ้อนของคำศัพท์ เนื่องจากศัพท์ควบคุมได้มีการกำหนดให้ใช้คำศัพท์เพียง 1 คำ แทนที่เนื้อหาของเอกสารเรื่องใดเรื่องหนึ่ง คำศัพท์อื่น ๆ ที่มีความหมายใกล้เคียงกัน ก็จะใช้วิธีโยงคำ ดังนั้นจึงเป็นการแก้ปัญหาเกี่ยวกับศัพท์ที่มีความหมายเหมือนกัน หรือใกล้เคียงกัน ส่วนคำศัพท์ที่มีมากกว่าหนึ่งความหมาย จะมีการอธิบายขอบเขตและความหมายของคำศัพท์นั้น ๆ การใช้ศัพท์ควบคุมทำให้คำศัพท์ที่ใช้นั้นเนื้อหาเอกสารมีความคงที่ และศัพท์ที่ผู้ทําการค้นใช้กับศัพท์ที่ผู้ค้นใช้ จะเป็นคำเดียวกัน
2. เพื่ออำนวยความสะดวกในการค้นคืนข้อมูล เนื่องจากศัพท์ควบคุมได้มีการรวมกลุ่มคำที่มีความหมายสัมพันธ์กัน เช่นอาจจะเรียงตามลำดับชั้นของคำที่มีความสัมพันธ์กัน ทำให้การค้นข้อมูลด้วยศัพท์ควบคุมทำได้อย่างกว้างขวางยิ่งขึ้น

การทําการค้นด้วยภาษาควบคุม

การทําการค้นที่ใช้ภาษาควบคุมต้องอาศัยสติปัญญาของมนุษย์ในการวิเคราะห์เนื้อหาของเอกสารเพื่อคัดเลือกเรื่อง หรือแนวความคิดที่จะนำมาทําการค้น แล้วคัดเลือกคำศัพท์จากภาษาธรรมชาติให้สอดคล้องกับเรื่องที่เลือกไว้แล้ว ดังนั้นผู้ทําการค้นจะต้องมีความเข้าใจหรือคุ้นเคยกับเนื้อหาในเอกสาร และรายการคำศัพท์ในภาษาธรรมชาติที่ใช้

การทําการค้นและภาษาของศัพท์ธรรมชาติ มีผลต่อการค้นคืนสารสนเทศ ปัจจัยสำคัญของการทําการค้นและภาษาธรรมชาติ ที่ส่งผลต่อประสิทธิภาพของการค้นคืนนั้น แบ่งออกได้ 4 ประการ คือ 1) นโยบายการทําการค้น 2) ภาษาของธรรมชาติ 3) ความถูกต้องในการทําการค้น และ 4) ความคงที่ในการทําการค้น (Pao, 1989: 125; Lancaster, 1979: 193)

นโยบายการทําการค้น สิ่งที่เกี่ยวข้องกับนโยบายของการทําการค้น ได้แก่ ความครอบคลุมกว้างขวางและละเอียดสมบูรณ์ (Exhaustivity) และปริมาณของศัพท์ธรรมชาติ (Density of Indexing) (Pao, 1989: 125)

1. ความครอบคลุมกว้างขวางและละเอียดสมบูรณ์ (Exhaustivity) ของการทำดัชนี หมายถึงจำนวนเรื่องหรือแนวความคิดในเอกสารที่นำมากำหนดดัชนี หรืออีกนัยหนึ่ง Exhaustivity นั้น วัดได้จากจำนวนของศัพท์ดัชนีในเอกสารแต่ละรายการ ระดับของ Exhaustivity เป็นผลจากการตัดสินใจด้านนโยบายของการทำดัชนี ในแต่ละห้องสมุดหรือแต่ละระบบ การทำดัชนีที่มี Exhaustivity สูง แสดงว่ามีการเลือกเรื่อง หรือแนวความคิดทุกประเด็นที่ปรากฏในเอกสารมากำหนดเป็นดัชนี เรียกการทำดัชนีในลักษณะนี้ว่า การทำดัชนีอย่างลึกซึ้ง (Depth Indexing หรือ Deep Indexing) ซึ่งมีเป้าหมายที่จะกำหนดดัชนีให้กับทุกประเด็นในเอกสาร ตรงกันข้ามกับการทำดัชนีแบบสรุป (Summarization) ที่เลือกทำดัชนีเฉพาะประเด็นสำคัญ ระดับของ Exhaustivity มีผลโดยตรงต่อการค้นคืน กล่าวคือ ถ้า Exhaustivity สูง ผลลัพธ์จากการค้นจะมีจำนวนสูง แต่ผลลัพธ์ที่ตรงกับความต้องการมีจำนวนต่ำ แต่ถ้า Exhaustivity ต่ำ ผลลัพธ์จะมีจำนวนต่ำ แต่มีผลลัพธ์ที่ตรงกับความต้องการสูงชัน (Lancaster, 1979: 194-195; Foskett, 1982: 25; Pao, 1989: 126-127)

2. ปริมาณของศัพท์ดัชนี (Density of Indexing) เป็นการวัดค่าเฉลี่ยของจำนวนศัพท์ดัชนีที่คัดเลือกเพื่อใช้แทนที่เนื้อหาของเอกสารแต่ละรายการ ปริมาณของศัพท์ดัชนีนี้ เกี่ยวข้องกับ Exhaustivity คือสามารถใช้ปริมาณของศัพท์ดัชนีไปประมาณค่าของ Exhaustivity ได้ เนื่องจากการวัดระดับของ Exhaustivity ทำได้ยาก ทั้งยังไม่สามารถสร้างเครื่องมือสำหรับวัดระดับของ Exhaustivity ได้ (Pao, 1989: 128)

ภาษาดัชนี คุณลักษณะของภาษาดัชนีที่เป็นปัจจัยต่อประสิทธิภาพของการค้นคืน คือ ความเฉพาะเจาะจง (Specificity)

ความเฉพาะเจาะจง (Specificity) ในการทำดัชนี หมายถึง ความสัมพันธ์แบบระดับชั้นของแนวความคิด (Generic relationship หรือ Genus/species relationship) เป็นความสัมพันธ์ระหว่างสิ่งของ และชนิดของสิ่งของ ตัวอย่างเช่น คำว่า Buildings ถือว่าเป็น Genus และ Houses เป็น Species ความสัมพันธ์ของ Buildings และ Houses จึงเป็นความสัมพันธ์แบบ Genus/species (Brown, 1982) Specificity เป็นคุณลักษณะของระบบดัชนีที่ยอมให้กำหนดศัพท์ดัชนีให้กับเนื้อหาของเอกสารได้ชัดเจนและเฉพาะเจาะจง ถ้าดัชนีมีความเฉพาะเจาะจงสูง จะทำให้ผลลัพธ์จากการค้นคืนตรงกับความต้องการสูงชันตามไปด้วย ในทางตรงกันข้าม ถ้าดัชนีไม่มีความเฉพาะเจาะจง ผลลัพธ์จากการค้นคืนจะมีจำนวนมาก แต่ที่ตรงกับความต้องการมีจำนวนน้อย (Foskett, 1982: 25)

ความถูกต้องในการทาดรรชนี นอกจากปัจจัยด้านนโยบายและภาษาของการทาดรรชนีแล้ว Lancaster (1979: 196) ยังได้กล่าวถึงความล้มเหลวในการค้นคว้าว่ามีสาเหตุมาจากการขาดความถูกต้องในการทาดรรชนี ซึ่งสามารถแบ่งออกได้ 2 ลักษณะ คือ

1. เรื่องที่เป็นประเด็นสำคัญในเอกสารไม่ได้รับการคัดเลือกมาทาดรรชนี ทำให้ผู้ใช้ไม่สามารถเข้าถึงเรื่องเหล่านั้นได้ จำนวนผลลัพธ์จากการค้นจึงต่ำกว่าความเป็นจริง
2. คำศัพท์ที่ใช้ไม่ตรงกับเรื่องในเอกสารหรือไม่มีความชัดเจนหรือเฉพาะเจาะจง ทำให้ผลลัพธ์จากการค้น และผลลัพธ์ที่ตรงกับความต้องการมีจำนวนน้อย

ความคงที่ในการทาดรรชนี (Consistency of Indexing) แบ่งออกได้ 2 ลักษณะคือ (Lancaster, 1979: 197)

1. ความคงที่ในการทาดรรชนีของผู้ทาดรรชนีหลายคน (Interindexer Consistency) หมายถึง การที่ผู้ทาดรรชนีตั้งแต่ 2 คน ขึ้นไปใช้คำศัพท์คำเดียวกันแทนเนื้อหาของเอกสารเรื่องใดเรื่องหนึ่ง
2. ความคงที่ในการทาดรรชนีของผู้ทาดรรชนีแต่ละคน (Intraindexer Consistency) หมายถึง การที่ผู้ทาดรรชนีคนหนึ่งใช้ศัพท์ดรรชนีคำเดียวกันทุก ๆ ครั้ง เพื่อแทนที่เนื้อหาของเอกสารเรื่องใดเรื่องหนึ่ง

การขาดความคงที่ในการทาดรรชนีทั้ง 2 กรณี จะทำให้จำนวนผลลัพธ์จากการค้น และผลลัพธ์ที่ตรงกับความต้องการคลาดเคลื่อนจากความเป็นจริง

ศัพท์อิสระ

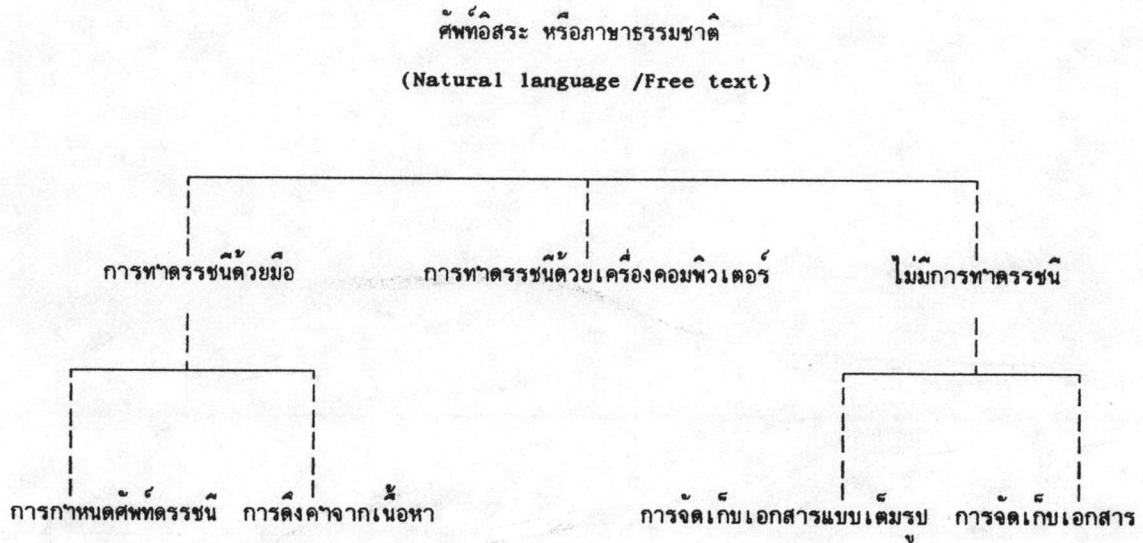
การทาดรรชนีที่ไม่มีการใช้ศัพท์ควบคุมถือว่าเป็นการทาดรรชนีแบบศัพท์อิสระหรือภาษาธรรมชาติ (Natural language) หรือบางครั้งเรียกว่า ระบบภาษาอิสระ (Free text system) (Lancaster, 1979: 279) การทาดรรชนีด้วยศัพท์อิสระ เป็นการกำหนดดรรชนี โดยใช้ภาษาของเอกสาร ส่วนใหญ่เป็นคำที่คอมพิวเตอร์กำหนดขึ้นจากชื่อเรื่องสาระสังเขป และเนื้อหาส่วนอื่น ๆ ของเอกสาร ในการคัดเลือกคำศัพท์ไม่จำเป็นต้องอ่านหรือวิเคราะห์เนื้อหาของเอกสาร (Rowley, 1988: 52-53)

ดรรชนีที่กำหนดด้วยศัพท์อิสระ เป็นคำที่มีความเฉพาะเจาะจงสูง และเป็นภาษาที่ใช้ในการสนทนาหรือสื่อสารกันในแต่ละสาขาวิชา เนื่องจากการจัดทาดรรชนีศัพท์อิสระใช้วิธีเลือกคำจากเอกสาร ไม่จำเป็นต้องอาศัยสติปัญญาของมนุษย์ ทำให้สามารถจัดทำได้อย่างรวดเร็ว และประหยัดค่าใช้จ่าย แต่ศัพท์อิสระมีปัญหาในด้านความซ้ำซ้อนของคำศัพท์ ทำให้เป็นภาระต่อผู้ค้นข้อมูลในการคิดคำศัพท์ที่จะใช้เป็นคำค้น

การทำคหรรชนีศัพท์อิสระ

Lancaster (1979: 279) ได้แสดงลักษณะของการทำคหรรชนีศัพท์อิสระ
ในรูปแผนภูมิ ดังนี้

แผนภูมิที่ 1 ลักษณะของการทำคหรรชนีศัพท์อิสระ



(ที่มา : Lancaster, 1979: 279)

จากแผนภูมิจะเห็นว่า การทำคหรรชนีภาษาธรรมชาติ หรือศัพท์อิสระนั้น สามารถทำได้โดยการทำได้ด้วยมือ (Human indexing) ทำด้วยเครื่องคอมพิวเตอร์ (Machine indexing) และไม่มีการทำคหรรชนี (No indexing) ในระบบที่ทำด้วยมือนั้น แบ่งออกได้ 2 วิธี คือ วิธีแรกเป็นการกำหนดศัพท์คหรรชนีขึ้นเอง (Assignment) โดยไม่ใช้คำจากรายการศัพท์ควบคุม และไม่จำเป็นต้องเป็นคำจากเนื้อหาของเอกสาร ส่วนวิธีที่ 2 เป็นการกำหนดศัพท์คหรรชนีโดยใช้คำจากเอกสาร ซึ่งอาจเป็นเนื้อหา หรือชื่อเรื่องเอกสารก็ได้ (Extraction) หากเป็นการทำคหรรชนีด้วยเครื่องคอมพิวเตอร์ หรือการทำให้ระบบอัตโนมัติ นั้น จะใช้วิธีการดึงคำจากเนื้อหาของเอกสารเพียงอย่างเดียว ส่วนในระบบที่ไม่มีการทำคหรรชนี แต่จะใช้คำทุกคำจากเอกสารเป็นคำค้น ในระบบนี้จะมีการจัดเก็บเอกสาร 2 ลักษณะ คือ การจัดเก็บเอกสารแบบเต็มรูป (Full text) และจัดเก็บเอกสารเพียงบางส่วน (Partial text) เช่น จัดเก็บในรูปแบบสาระสังเขป เป็นต้น

การใช้ศัพท์อิสระในการค้นคืนสารสนเทศ เกิดขึ้นเป็นครั้งแรกใน ค.ศ. 1951 เมื่อ Taube ได้คิดระบบ Uniterm ซึ่งมีหลักการสำคัญ คือ เรื่องหรือเนื้อหาของเอกสารจะแทนที่ได้ด้วยคำศัพท์คำเดียว (Uniterms) ซึ่งเป็นคำที่ดึงมาจากเนื้อหาของเอกสาร ผู้ทราตรรชนีจึงไม่จำเป็นต้องมีความรู้ ความชำนาญมากนัก อย่างไรก็ตามระบบนี้มีปัญหาหลายประการ ปัญหาบางประการเป็นเรื่องที่ได้รับการแก้ไขมาแล้วในระบบศัพท์ควบคุม ด้วยเหตุนี้จึงมีการกลับไปใช้ศัพท์ควบคุมและนำไปสู่การพัฒนาาระบบศัพท์สัมพันธ์ในเวลาต่อมา (Lancaster, 1989: 2-3)

นอกเหนือจากปัญหาด้านศัพท์แล้ว ระบบ Uniterm ยังมีข้อจำกัดเรื่องการค้นข้อมูล เนื่องจากในการค้นข้อมูลเรื่องใดเรื่องหนึ่งนั้น จะต้องเปรียบเทียบคำตรรชนีจากบัตรตั้งแต่ 2 บัตร ขึ้นไป ซึ่งสร้างความยุ่งยากในการค้นแก่ผู้ใช้ ต่อมาจึงได้นำคอมพิวเตอร์มาใช้เพื่อแก้ปัญหาดังกล่าว

การทราตรรชนีระบบอัตโนมัติ

การทราตรรชนีในระบบที่ใช้คนทำ ประกอบด้วย 2 ขั้นตอน คือการวิเคราะห์เนื้อหา และการคัดเลือกคำศัพท์ ซึ่งต้องใช้สติปัญญาของมนุษย์ ทำให้มีค่าใช้จ่ายสูงและสิ้นเปลืองเวลา จึงเกิดความพยายามที่จะใช้คอมพิวเตอร์มาช่วยเพื่อลดภาระงานดังกล่าว การใช้คอมพิวเตอร์ในการทราตรรชนีไม่ว่าในขั้นตอนใดขั้นตอนหนึ่ง หรือทั้งสองขั้นตอน ถือว่าเป็นการทราตรรชนีระบบอัตโนมัติ (Pao, 1989: 131)

การนำคอมพิวเตอร์มาใช้ในการค้นคืนข้อมูลอย่างจริงจัง เริ่มขึ้นเมื่อปลายทศวรรษ 1950 และต้นทศวรรษ 1960 (Lancaster, 1989: 3) ในที่นี้จะกล่าวถึงการทราตรรชนีระบบอัตโนมัติประเภทต่าง ๆ ตามลักษณะการสร้างคำตรรชนีของเครื่องคอมพิวเตอร์ ดังนี้

การทราตรรชนีจากชื่อเรื่อง

โดยทั่วไปผู้แต่งมักจะกำหนดชื่อเรื่องให้สะท้อนเนื้อหาของหนังสือ ชื่อเรื่องจึงแสดงให้ผู้อ่านทราบว่าหนังสือเป็นเรื่องที่เกี่ยวกับอะไร หากเป็นหนังสือที่มีเนื้อหาประเภทเดียวกัน แต่แต่งโดยผู้แต่งคนละคน ผู้แต่งแต่ละคนมักจะตั้งชื่อหนังสือให้แตกต่างกัน แต่จะมีคำสำคัญหรือคำหลักคำเดียวกัน (Foskett, 1982: 39) ตัวอย่างเช่น

Manual of library classification
Library classification on the march
Introduction to library classification
Prolegomena to library classification

จะเห็นว่าคำสำคัญในแต่ละเรื่องเป็นคำเดียวกัน คือ Library classification ดังนั้น คำศัพท์จากชื่อเรื่องของเอกสารจึงเป็นสิ่งที่ช่วยให้ผู้ใช้เข้าถึงเอกสารที่เกี่ยวข้องได้ การใช้คำสำคัญจากชื่อเรื่องมาเป็นคำค้นหรือศัพท์ดรรรชนีนั้นมีหลายวิธี แต่วิธีการที่จะกล่าวถึงในที่นี้คือ การจัดเรียงหรือหมุนเวียนคำศัพท์ในชื่อเรื่องเพื่อให้ทำหน้าที่เป็นศัพท์ดรรรชนี ซึ่งส่วนใหญ่จะจัดทำโดยใช้เครื่องคอมพิวเตอร์ ตัวอย่างการทาดรรรชนีในลักษณะนี้ได้แก่

การทาดรรรชนีโดยการดึงคำจากชื่อเรื่อง (Catching Title indexing)

Catching Title เป็นการทาดรรรชนีโดยการหมุนเวียนคำในชื่อเรื่องให้อยู่ที่ตำแหน่งแรกของชื่อเรื่อง ซึ่งคำที่ปรากฏที่ตำแหน่งแรกของชื่อเรื่องนี้จะ เป็นศัพท์ดรรรชนี ด้วยวิธีการดังกล่าว จึงสามารถกำหนดให้ทุกคำในชื่อเรื่องเป็นดรรรชนีได้ แต่เพื่อให้ศัพท์ดรรรชนีเป็นคำที่สื่อความหมายของเนื้อหาเอกสารได้อย่างสมบูรณ์ จึงต้องมีการกำหนดหลักเกณฑ์ของการหมุนเวียนคำ ตัวอย่างการหมุนเวียนคำศัพท์อย่างมีหลักเกณฑ์ จากชื่อเรื่อง "Field plotting by Fourier synthesis digital computation (Foskett, 1982: 39) มีดังนี้

Field plotting by Fourier synthesis and digital computation
Plotting, field, by Fourier synthesis and digital computation
Fourier synthesis and digital computaion, field plotting by
Synthesis, Fourier, and digital computation, field plotting by
Digital computation and Fourier synthesis, field plotting by
Computation, digital, and Fourier synthesis, field lotting by

จะเห็นได้ว่าการหมุนเวียนคำแบบมีหลักเกณฑ์นี้ จะต้องมีคำอื่นในชื่อเรื่องมาขยายคำดรรรชนี โดยคำที่มาขยายจะอยู่หลังเครื่องหมาย ", " เช่น Potting, field ดังนั้นในการจัดทำจึงต้องอาศัยผู้ที่มีความรู้ด้านการทาดรรรชนี แต่หากเป็นการหมุนเวียนคำศัพท์ที่ไม่มีกำหนดหลักเกณฑ์ไว้ล่วงหน้า จะได้ดรรรชนีในลักษณะดังต่อไปนี้ (Foskett, 1982: 39)

Field plotting by Fourier synthesis and digital computation
Plotting, by Fourier synthesis and digital computation, Field
Fourier synthesis and digital computaion, Field plotting by
Synthesis, and digital computation, field plotting by Fourier
Digital computation Field plotting by Fourier synthesis and
Computation, Field plotting by Fourier synthesis and digital.

การหมุนเวียนแบบไม่มีหลักเกณฑ์นี้ จะเห็นว่า คำ "Field plotting" และ "Fourier synthesis" นั้น เมื่อคำหลังคือ "plotting" และ "Synthesis"

เป็นคำค้น จะไม่สื่อความหมายได้อย่างสมบูรณ์ เท่ากับคำค้นที่มีการควบคุมการหมุนเวียนคำศัพท์ อย่างเช่น "Plotting, field" และ "Synthesis, Fourier"

การทำดัชนีแบบ KWIC

KWIC (Keyword in context) เป็นการทำดัชนีโดยใช้คำจากชื่อเรื่องโดยวิธีการหมุนเวียนคำ เช่นเดียวกับดัชนีแบบ Catching Title แต่ต่างกันที่ดัชนีแบบ KWIC นั้น คำศัพท์จากชื่อเรื่องที่เป็นดัชนีจะอยู่ตรงกลาง ข้อมูลชื่อเรื่องที่ไม่ได้เป็นศัพท์ดัชนีจะอยู่ด้านใดด้านหนึ่งของดัชนี ตัวอย่างการทำดัชนีแบบ KWIC จากชื่อเรื่องเดิมคือ Field Plotting by Fourier synthesis and digital computation (Foskett, 1982: 40) มีดังนี้

digital computation/ Field plotting by Fourier synthesis and
Computation/Field plotting by Fourier synthesis and digital
Field plotting by Fourier synthesis and digital computation/
plotting by Fourier synthesis and digital computation/Field
Fourier synthesis and digital computation/Field plotting by
synthesis and digital computation/Field plotting by Fourier.

จะเห็นว่าศัพท์ดัชนีจากการหมุนเวียนคำในชื่อเรื่องโดยเครื่องคอมพิวเตอร์นั้น เป็นคำที่สื่อเนื้อหาของเอกสาร ส่วนคำบางคำ เช่น and และ by จะไม่ได้รับคัดเลือกเป็นดัชนี ทั้งนี้เพราะการทำดัชนีแบบ KWIC นั้น จะสร้างดัชนีจากคำทุกคำในชื่อเรื่องของเอกสาร ยกเว้น คำที่อยู่ในรายการคำยกเว้น (Stop list หรือ Stop wordlist) ซึ่งเป็นรายการคำที่ไม่ได้สื่อเนื้อหาของเอกสาร แต่ปรากฏในชื่อเรื่องเพื่อทำหน้าที่เป็นคำเชื่อม หรือคำขยาย คำเหล่านี้ ได้แก่ a, an, the, how, it, whether, where เป็นต้น นอกจากนี้อาจมีคำศัพท์อื่น ๆ ที่มีได้ เป็นคำเชื่อมหรือคำขยาย แต่เป็นคำที่ไม่ประสงค์ให้ใช้เป็นคำค้น คำในรายการ stop wordlist นั้น แม้จะไม่ปรากฏเป็นคำค้นแต่จะปรากฏร่วมกับข้อความส่วนที่เหลือของชื่อเรื่อง ศัพท์ดัชนีจะเรียงตามลำดับอักษร ข้อมูลที่ให้รายละเอียดเกี่ยวกับแหล่งของเอกสารนั้น หากมีเนื้อที่เพียงพอจะมีรายละเอียดเกี่ยวกับรายการทางบรรณานุกรม ซึ่งประกอบด้วย ชื่อวารสาร ปีที่ ฉบับที่ และเลขหน้า ส่วนรายละเอียดทางบรรณานุกรมที่สมบูรณ์นั้นจะอยู่ในรายการซึ่งแยกต่างหาก

การทำดัชนีแบบ KWOC

KWOC (Keyword-out-of-context) เป็นการทำดัชนีจากชื่อเรื่อง เช่นเดียวกับการทำดัชนีแบบ KWIC แต่ต่างกันที่คำสำคัญจะปรากฏอยู่ทางซ้ายมือ ตามหลังด้วยชื่อเรื่องที่สมบูรณ์ ดัชนีแบบ KWOC นี้ได้รับการพัฒนาเพื่อแก้ปัญหาความไม่คุ้นเคยของผู้ใช้ที่มีต่อ

ดรชชนนี้แบบ KWIC ที่ศัพท์ดรชชนนี้ปรากฏอยู่ตรงกลางของชื่อเรื่อง ตัวอย่างของการทำดรชชนนี้แบบ KWOC (Foskett, 1982: 43) มีดังนี้

Library A modern outline of library classification
 Library Introduction to library classification
 Library Library classification on the march
 Library Library education
 Library Manual of library classification
 Library Mechanized acquisition procedures in the
 University of

การทำดรชชนนี้แบบ KWOC นั้น ศัพท์ดรชชนนี้อาจจะอยู่คนละบรรทัดกับชื่อเรื่อง และหมายเลขของเอกสาร และยังมีการใช้เครื่องหมาย "*" แทนคำในชื่อเรื่องที่ได้รับการคัดเลือกเป็นศัพท์ดรชชนนี้ ดังตัวอย่าง (Rowley, 1988: 96)

ADULTS

A DESCRIPTIVE STUDY OF BLACK AMERICAN * OF ACHIEVEMENT FROM
 AN EXISTENTIAL FRAMEWORK 097861

A STUDY OF THE VISUAL LANGUAGE PROCESSING ABILITIES OF
 BRAIN INJURED * APRAXIC VERSUS DYSARTHIC SUBJECTS 097865

VISUAL SEQUENTIAL RECALL OF ASSOCIATIVE AND NON-ASSOCIATIVE
 STIMULI IN UNILATERALLY BRAIN-DAMAGED AND NORMAL * 067524

ADVANCED

THE * EVOLUTION OF GLOBULAR CLUSTER STARS ADVANTAGE 045879

THE RIGHT EAR * FOR THE PROCESSING OF LINGUISTIC STIMULI

067890

การทำดรชชนนี้จากเนื้อหาของเอกสาร

การกำหนดดรชชนนี้จากเนื้อหาของเอกสารโดยเครื่องคอมพิวเตอร์ เป็นวิธีการทำดรชชนนี้ที่มีการจัดเก็บเนื้อหาของเอกสารไว้ในเครื่องคอมพิวเตอร์ การเข้าถึงเอกสารดังกล่าวจำเป็นต้องมีการจัดทำดรชชนนี้ เพื่อใช้เป็นเครื่องมือในการค้นคืน การที่จะให้เครื่องคอมพิวเตอร์กำหนดดรชชนนี้จากคำทุกคำในเนื้อหาเอกสาร เหมือนกับการกำหนดดรชชนนี้จากคำในชื่อเรื่องนั้น ทำให้ต้องใช้เนื้อที่ในการจัดเก็บข้อมูลดรชชนนี้สูงมาก จึงต้องให้เครื่องคอมพิวเตอร์คัดเลือกเฉพาะคำสำคัญในเอกสารมาทำดรชชนนี้ ในการพิจารณาว่าคำใดเป็นคำสำคัญนั้น เป็นกิจกรรมที่ต้องอาศัยสติปัญญาและวิจารณญาณของมนุษย์ แต่หากจำเป็นต้องใช้เครื่องคอมพิวเตอร์ ก็จะต้องมีการกำหนดวิธีการจัดทำเป็นพิเศษ

วิธีการขั้นพื้นฐานสำหรับการจัดทราจรรชนีจากเนื้อหาของเอกสารด้วยเครื่องคอมพิวเตอร์ คือการนับความถี่ของคำที่ปรากฏในเอกสาร (Pao, 1989: 134) เนื่องจากคำที่ปรากฏในเอกสารบ่อยครั้ง จะเป็นคำที่มีความสัมพันธ์กับเนื้อหาของเอกสาร วิธีการนับความถี่ของคำที่มีประสิทธิภาพ คือการนับจำนวนครั้งที่คำนั้น ๆ ปรากฏในเอกสาร แล้วนำไปเปรียบเทียบกับเกณฑ์มาตรฐาน ซึ่งประกอบด้วยรายการคำ และจำนวนความถี่ หากคำศัพท์คำใดในเอกสารมีค่าความถี่สูงกว่าเกณฑ์มาตรฐาน ถือว่าคำนั้นมีความสำคัญ สามารถกำหนดเป็นศัพท์ดรรชนีได้ (Foskett, 1982: 49)

เทคนิคการกำหนดคำศัพท์จากเนื้อหาเอกสารอีกรูปแบบหนึ่ง คือการเลือกคำจากรายการคำ ที่เรียงตามลำดับของความถี่จากมากไปหาน้อย โดยเลือกเฉพาะคำศัพท์ที่อยู่ในอันดับต้น ๆ เป็นจำนวนร้อยละ 5 ของจำนวนคำทั้งหมดในรายการคำ ทั้งนี้จะต้องตัดคานำหน้านาม คำบุพบท คำเชื่อม หรือคำอื่น ๆ ที่ไม่มีความหมาย ซึ่งปรากฏอยู่ใน Stop list (Pao, 1989: 134)

วิธีการเลือกคำสำคัญจากเอกสารอีกวิธีหนึ่ง คือเลือกเฉพาะคำที่มีเนื้อหาสาระ และจะต้องปรากฏในย่อหน้าหนึ่ง ๆ จำนวน 2 ครั้งขึ้นไป (Pao, 1989: 134)

การทราจรรชนีภาษาไทยระบบอัตโนมัติ

การกำหนดดรรชนีจากคำศัพท์ในชื่อเรื่อง เป็นความสามารถประการหนึ่งของซอฟต์แวร์ระบบงานห้องสมุดที่มีจำหน่ายอยู่ในปัจจุบัน โปรแกรมทางด้านห้องสมุดที่มีใช้ในห้องสมุดในประเทศไทย เช่น Micro CDS/ISIS, BRS หรือซอฟต์แวร์แบบบูรณาการ เช่น DYNIX ล้วนมีคุณลักษณะที่สามารถสร้างดรรชนีจากชื่อเรื่อง และเข้าถึงข้อมูลได้ด้วยคำสำคัญจากชื่อเรื่องเช่นเดียวกัน

การทราจรรชนีจากคำในชื่อเรื่องจึงทำได้สะดวก รวดเร็ว และไม่ต้องมีคนเข้าไปเกี่ยวข้องในระหว่างการจัดทำ ผู้ใช้เพียงแต่กำหนดความต้องการเกี่ยวกับการทราจรรชนีในขั้นตอนการกำหนดโครงสร้างของข้อมูลเท่านั้น

อย่างไรก็ตามซอฟต์แวร์ดังกล่าวได้รับการพัฒนาเพื่อรองรับการประมวลผลข้อมูลที่เป็นภาษาอังกฤษหรือภาษาอื่นที่ใช้อักษรโรมัน เมื่อนำมาใช้ในการประมวลผลข้อมูลที่เป็นภาษาไทย จึงมีปัญหามากหลายประการ

ปัญหาประการแรก คือ โครงสร้างภาษาไทยมีเอกลักษณ์ที่แตกต่างจากโครงสร้างภาษาอังกฤษ กล่าวคือภาษาอังกฤษมีอักขระระดับเดียว แต่ตำแหน่งของอักขระจะมีอักขระเพียงตัวเดียว และอ่านจากซ้ายไปขวา (ลัดดา ยินดีมาก โกรติ, 2533: 1) ลักษณะเช่นนี้จะเป็นปัญหาต่อการแสดงผล และการจัดเรียงคำตามลำดับอักษรของคำที่ขึ้นต้นด้วยสระ และคำที่มีทั้งสระและ

วรรณยุกต์อยู่ในตำแหน่งเดียวกัน แต่ปัญหาดังกล่าวได้รับการแก้ไขโดยการพัฒนาโปรแกรมเพิ่มเติม เพื่อให้สามารถประมวลผลภาษาไทยได้ถูกต้องมากที่สุด

ปัญหาอีกประการหนึ่ง คือไม่มีการเว้นวรรคระหว่างคำในภาษาไทย (ลัดดา ยินดีมาก โกรติ, 2533: 1) เครื่องคอมพิวเตอร์จึงไม่สามารถดึงคำจากข้อเรื่องออกมาทำคหรรชนีได้ เนื่องจากโปรแกรมได้รับการออกแบบเพื่อประมวลผลภาษาอังกฤษดังได้กล่าวแล้ว และภาษาอังกฤษมีการเว้นวรรคระหว่างคำ คอมพิวเตอร์จึงสามารถดึงคำในข้อเรื่องมาเป็นศัพท์คหรรชนีได้

ดังนั้นการทำคหรรชนีจากข้อเรื่องภาษาไทยจึงต้องแบ่งคำในข้อเรื่องออกเป็นคำ ๆ อย่างไม่รู้ก็ตาม การแบ่งคำในภาษาไทยก็ยังคงเป็นปัญหา เนื่องจากมีอาจกระทำได้โดยอาศัยหลักไวยากรณ์เท่านั้น แต่จะต้องอาศัยความหมายของคำในประโยคนั้น ๆ ด้วย (ลัดดา ยินดีมาก โกรติ, 2533: 1) ปัญหาเรื่องการแบ่งคำจึงเป็นสาเหตุประการหนึ่งที่ทำให้มีการทำคหรรชนีภาษาไทยตามลักษณะภาษาธรรมชาติน้อยมาก

ลัดดา ยินดีมาก โกรติ (2533) ได้ศึกษาเกี่ยวกับการประมวลผลภาษาไทยตามลักษณะธรรมชาติ เรื่อง การทำคหรรชนีคำสำคัญอัตโนมัติ โดยการทดลองใช้เทคนิค ซึ่งใช้กับภาษาอังกฤษ ผสมผสานกับเทคนิคการหาความถี่ของคำที่อยู่ติดกัน ซึ่งมีวิธีการโดยสรุปดังนี้

1. ตัดคำออกเป็นคำย่อย ๆ ที่สามารถอ่านได้ หรือเรียกว่า Token (อาจมีพยางค์เดียวหรือมากกว่าก็ได้)
2. ตัด Token ที่ไม่สำคัญหรือคายกเว้น (Stopword) ออก
3. คำนวณหาความถี่ของ Token
4. นำ Token ที่มีความถี่สูงมากกำหนดค่า Token ที่อยู่ติดกันเพื่อประกอบเป็นคำ หรือวลี
5. นำคำหรือวลีมาให้น้ำหนัก เพื่อเลือกคำที่มีน้ำหนักมากเป็นคำสำคัญ
6. กำหนดจำนวนคำต่อเอกสาร

จากการประเมินผลคำสำคัญอัตโนมัติที่ได้จากการทดลองพบว่า มีค่า Recall เท่ากับ .625 และค่า Precision เท่ากับ .537 ซึ่งผู้วิจัยคิดว่าผลที่ได้เป็นที่น่าพอใจ

ส่วนการแบ่งคำในระบบมีอนั้น ยังคงเป็นวิธีเดียวที่ช่วยให้เครื่องคอมพิวเตอร์สามารถทำคหรรชนีศัพท์อิสระได้ เพราะซอฟต์แวร์ห้องสมุด ไม่ว่า BRS, DYNIX หรือ VTLs หากต้องการให้เครื่องทำคหรรชนีโดยใช้คำในเขตข้อมูลใด ๆ นั้น จะต้องมีการเว้นวรรคระหว่างคำจำนวน 1 ระยะ ซึ่งแบ่งคำในภาษาไทยนั้นยังคงมีปัญหาดังได้กล่าวแล้ว

สำหรับโปรแกรม Micro CDS/ISIS Version 3.0 นั้น เป็นโปรแกรมซึ่งผู้ใช้สามารถเลือกเขตข้อมูลที่ต้องการนำมาทำดัชนีได้ตามความพึงพอใจ ส่วนวิธีการที่เครื่องคอมพิวเตอร์ตั้งค้ำัพท์มาทำดัชนีนั้น มี 9 วิธี หรือ 9 เทคนิค ดังนี้

- 1) เทคนิค 0 หมายถึง การกำหนดให้ข้อมูลทั้งเขตข้อมูลเป็นคำค้น
- 2) เทคนิค 1,5 หมายถึง การกำหนดให้ข้อมูลทุกเขตข้อมูลย่อยเป็นคำค้น
- 3) เทคนิค 2,6 หมายถึง การกำหนดให้ข้อมูลที่อยู่ในเครื่องหมาย < > เป็นคำค้น
- 4) เทคนิค 3,7 หมายถึง การกำหนดให้ข้อมูลที่อยู่ในเครื่องหมาย / / เป็นคำค้น
- 5) เทคนิค 4,8 หมายถึง การกำหนดให้คำทุกคำในเขตข้อมูลที่ระบุ เป็นคำค้น

ทั้งนี้ เทคนิค 5-8 สามารถเติมข้อความนำหน้าศัพท์ดัชนีได้ เช่น เพิ่มคำว่า "TI=" เป็นต้น อย่างไรก็ตาม ในการกำหนดดัชนีตามเทคนิคที่ 2-4 และ 6-8 (ข้อ 3-5) จะต้องมีการคัดเลือก หรือแบ่งคำที่ต้องการทำดัชนี ด้วยระบบมือเช่นเดียวกัน

ในการคัดเลือกศัพท์เพื่อกำหนดเป็นดัชนีนั้น หากต้องการใช้หลักการแบ่งคำตามความหมายของคำในเอกสาร นววรรณ พันธุมธา (2527) ได้จำแนกคำตามหน้าที่ในการสื่อสารออกเป็น 6 หมวด หมวดคำซึ่งเป็นคำสำคัญที่ผู้พูดใช้บอกเนื้อความซึ่งสื่อสารไปยังผู้ฟัง คือ คำหลัก คำหลักนี้ แบ่งย่อยตามความหมายได้ 2 ประเภท คือ คำนามและคำกริยา ส่วนวลี หมายถึง กลุ่มคำซึ่งเกิดจากการนำคำมาขยายคำหลัก ในเมื่อคำหลักมี 2 ชนิด คือ คำนาม และ คำกริยา วลีจึงมี 2 ชนิด เช่นเดียวกัน คือ นามวลีและกริยาวลี ดังนั้น เพื่อให้การกำหนดดัชนีประเภทศัพท์อิสระ เป็นไปตามหลักไวยากรณ์ภาษาไทย ดรรชนีประเภทศัพท์อิสระจึงควรมีลักษณะตามข้อหนึ่งข้อใดดังนี้

1. เป็นคำนามสามัญ คือ คำที่หมายถึงสิ่งต่าง ๆ ทั่วไป เช่น การศึกษา นักเรียน
2. เป็นคำนามวิสามัญ คือ คำที่หมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ เช่น กรุงเทพมหานคร เช่น ญุ่น
3. เป็นคำกริยา คือ คำที่ใช้แสดงอาการ สภาพหรือเหตุการณ์ต่าง ๆ เช่น พัฒนา วัตถุประสงค์
4. เป็นนามวลี คือ กลุ่มคำซึ่งประกอบด้วยคำนาม ซึ่งเป็นคำหลักและคำอื่น ๆ เป็นคำขยาย เช่น ศูนย์วิชาการ แผนการสอน
5. เป็นกริยาวลี คือ กลุ่มคำซึ่งประกอบด้วยคำกริยาซึ่งเป็นคำหลักและคำอื่น ๆ เป็นคำขยาย เช่น รัฐบาล เกษียณอายุราชการ
6. เป็นคำทับศัพท์ภาษาต่างประเทศใช้ตามที่ปรากฏในชื่อบทความ เช่น วิดีโอ วาลดอร์ฟ

การวัดประสิทธิภาพของครรชน

ความสำเร็จของบริการสารนิเทศพิจารณาได้จากผลการค้นข้อมูลของผู้ใช้บริการว่าตรงกับความต้องการหรือไม่และข้อมูลที่ค้นได้มีความถูกต้องสมบูรณ์เพียงใด การวัดความถูกต้องสมบูรณ์ของผลการค้นเป็นสิ่งจำเป็นต่อการประเมินการค้นคืนสารนิเทศ เครื่องมือที่นิยมใช้ประเมินประสิทธิภาพของการค้นคืน คือ Recall และ Precision (Lancaster, 1979: 111)

Recall (ข้อมูลเรื่องใดเรื่องหนึ่งซึ่งค้นคืนได้)

Recall หมายถึง สัดส่วนของข้อมูลเรื่องใดเรื่องหนึ่งในฐานข้อมูลที่ใช้ค้นคืนได้ (Lancaster, 1991: 285) Recall เป็นเครื่องมือสำหรับวัดว่าข้อมูลในเรื่องที่ต้องการสามารถค้นคืนได้หรือไม่เพียงใด ในการค้นคืนข้อมูลนั้น ผลการค้นแม้มีทั้งรายการที่ตรงและไม่ตรงกับความต้องการ หากผู้ค้นต้องการให้ผลการค้นครบถ้วนสมบูรณ์ สิ่งซึ่งจะบ่งบอกถึงความครบถ้วนสมบูรณ์ คือจำนวนข้อมูลที่เกี่ยวข้องสามารถค้นออกมาได้ การวัดความสมบูรณ์ของการค้นฐานข้อมูล จะเรียกว่า Recall ratio (อัตราส่วนของข้อมูลเรื่องใดเรื่องหนึ่งที่ค้นคืนได้) ถ้ากล่าวว่ recall ร้อยละ 80 หมายถึง ในฐานข้อมูลมีข้อมูลในเรื่องนั้น ๆ รวม 10 รายการ แต่ค้นคืนได้เพียง 8 รายการ จากตารางที่ 1 Recall ratio สามารถแสดงในรูปสูตรได้ดังนี้

$$\text{Recall ratio (อัตราส่วนของข้อมูลเรื่องใดเรื่องหนึ่งที่ค้นคืนได้)} = \frac{\text{ข้อมูลที่ค้นคืนได้และเกี่ยวข้อง}}{\text{ข้อมูลที่เกี่ยวข้องทั้งหมด}} \quad \text{หรือ} \quad \frac{a}{a+c}$$

ตารางที่ 1 แสดงผลจากการค้นคืนข้อมูล

	การพิจารณาความเกี่ยวข้อง		
	เกี่ยวข้อง	ไม่เกี่ยวข้อง	รวม
ค้นคืนได้	a (ตรง)	b (ไม่ตรง)	a+b
ค้นคืนไม่ได้	c (ขาดหายไป)	d (ไม่ต้องการ)	c+d
รวม	a+c	b+d	a+b+c+d (ข้อมูลทั้งหมดในฐานข้อมูล)

(ที่มา : Lancaster, 1979: 113)

Precision (ความตรงกับความต้องการของข้อมูลที่ค้นได้)

Precision เป็นอัตราส่วนของรายการที่ค้นคืนได้ซึ่งผู้ใช้เห็นว่าเกี่ยวข้องกับเรื่องที่ต้องการ ตัวอย่างเช่น การค้นข้อมูลซึ่งได้ผลการค้น 50 รายการ ในจำนวนนี้มีข้อมูลที่ผู้ใช้เห็นว่าตรงกับความต้องการ จำนวน 10 รายการ ดังนั้น Precision Ratio (อัตราส่วนของความตรงของข้อมูลที่ค้นได้) จึงมีค่าเท่ากับ 10/50 จากตารางที่ 1 อัตราส่วนของ Precision แสดงในรูปสูตร ได้ดังนี้

$$\text{Precision Ratio (อัตราส่วนของความตรงของข้อมูลที่ค้นได้)} = \frac{\text{จำนวนข้อมูลที่ค้นได้และเกี่ยวข้อง}}{\text{จำนวนข้อมูลที่ค้นคืนได้}} \text{ หรือ } \frac{a}{a+b}$$

Recall เป็นเครื่องมือสำหรับวัดความสามารถของระบบในการค้นคืนข้อมูลที่เกี่ยวข้อง ในขณะที่ Precision นั้นเป็นการวัดความสามารถของระบบในการคัดแยกสิ่งที่ไม่เกี่ยวข้องออกไป ในการพิจารณาประสิทธิภาพของระบบ ว่าสามารถดึงข้อมูลที่ตรงกับความต้องการ และแยกสิ่งที่ไม่เกี่ยวข้องออกไปได้ดีเพียงใดนั้น ต้องใช้ทั้ง Recall และ Precision ควบคู่กันไป การใช้ Recall หรือ Precision เพียงอย่างเดียวอย่างหนึ่งจะไม่สามารถแสดงถึงประสิทธิภาพของระบบได้อย่างสมบูรณ์ ถึงแม้ในบางครั้งจำนวน Recall จะสูงถึงร้อยละ 100 (คือค้นข้อมูลได้ทั้งหมด ซึ่งเท่ากับ a+b+c+d ในตารางที่ 1) แต่จำนวน Precision ก็จะต่ำลงเนื่องจาก Recall และ Precision มีความสัมพันธ์แบบผกผัน (Lancaster, 1979: 113) กล่าวคือ ผลการค้นที่มี Recall สูง Precision จะต่ำ ในทางตรงกันข้ามเมื่อมีการจำกัดขอบเขตการค้นเพื่อให้มี Precision สูงขึ้น จำนวน Recall จะต่ำลง

ความต้องการเกี่ยวกับจำนวน Recall และ Precision ของผู้ค้นแต่ละคนจะแตกต่างกัน ขึ้นอยู่กับวัตถุประสงค์ของการค้น หากต้องการสำรวจข้อมูลในเบื้องต้นเพื่อการวิจัย ผลการค้นจะต้องมี Recall สูง แต่หากต้องการข้อมูลเฉพาะเรื่อง หรือข้อเท็จจริงในเรื่องใดเรื่องหนึ่ง ผลการค้นจะต้องมี Precision สูง

การเปรียบเทียบคุณลักษณะของศัพท์ควบคุมและศัพท์อิสระ

ดรธรชนีประเภทศัพท์ควบคุม และศัพท์อิสระต่างมีข้อดี และข้อด้อยทั้งในด้านการจัดทำและการค้นข้อมูล (Harter, 1986: 51) โดยศัพท์ดรธรชนีทั้ง 2 ประเภทนี้ มีคุณลักษณะที่แตกต่างในด้านภาษา และด้านค่าใช้จ่าย ดังนี้

ในด้านภาษา ดรธรชนีประเภทศัพท์ควบคุมมีข้อดี คือ มีการควบคุมคำศัพท์ที่มีความหมายเหมือนกันหรือใกล้เคียงกัน โดยกำหนดให้ใช้ศัพท์ที่เหมาะสมเพียงคำเดียวแทนเนื้อหาในเรื่องใดเรื่องหนึ่งและมีการโยงคำอื่น ๆ ที่มีความหมายเหมือนกันไปยังคำที่กำหนดให้ใช้ นอกจากนี้ศัพท์ควบคุมประเภทศัพท์สัมพันธ์จะมีการรวบรวมคำที่เนื้อหาเกี่ยวข้องกันมาไว้ด้วยกัน แล้วระบุความ

สัมพันธ์ระหว่างกันว่ามีความหมายกว้างกว่าหรือแคบกว่า เป็นต้น หากเป็นคำ ๆ เดียวกันแต่มีความหมายมากกว่า 1 อย่าง ก็จะมีคำอธิบายเพิ่มเติม ดังนั้นศัพท์ควบคุมจึงอำนวยความสะดวกต่อการค้น เนื่องจากผู้ใช้ไม่ต้องคิดคำศัพท์หลายคำที่มีความหมายเหมือนกัน

ส่วนธรรมชาติประเภทศัพท์อิสระ หรือศัพท์ไม่ควบคุมนั้น มีลักษณะเด่น คือ คำศัพท์มีความเฉพาะเจาะจงสูง และเป็นคำที่กำหนดขึ้นโดยผู้แต่ง จึงสามารถสะท้อนเนื้อหาของเอกสารได้เป็นอย่างดี นอกจากนี้ศัพท์อิสระเป็นคำที่ใช้ติดต่อกันในสาขาวิชานั้น ๆ การค้นข้อมูลด้วยภาษาที่ผู้ใช้มีความคุ้นเคย ทำให้ค้นข้อมูลได้อย่างสะดวกมากขึ้น

ข้อด้อยในด้านภาษาของศัพท์ควบคุม คือคำศัพท์ไม่มีความเฉพาะเจาะจง ถึงแม้ว่าได้มีการพัฒนาระบบศัพท์สัมพันธ์เพื่อให้มีความเฉพาะเจาะจงมากกว่ารายการหัวเรื่อง แต่ก็ยังมีความเฉพาะเจาะจงน้อยกว่าศัพท์อิสระ นอกจากนี้ ศัพท์ควบคุมเป็นชุดของคำที่กำหนดขึ้นโดยคนกลุ่มหนึ่ง และมีคนอีกกลุ่มหนึ่งที่หน้าทีคัดเลือก เพื่อใช้แทนเนื้อหาของเอกสาร ดังนั้นการใช้ศัพท์ควบคุมได้อย่างถูกต้อง และมีประสิทธิภาพนั้นจะต้องใช้เวลาในการศึกษา และทำความเข้าใจกับคำศัพท์

ส่วนศัพท์อิสระมีข้อด้อย คือ คำศัพท์ไม่เป็นมาตรฐาน คลุมเครือ และคำศัพท์ที่แทนเรื่องใดเรื่องหนึ่งไม่มีความคงที่

ในด้านค่าใช้จ่ายนั้น การทำดัชนีด้วยศัพท์ควบคุมมีค่าใช้จ่ายสูง เพราะเป็นกิจกรรมที่ต้องอาศัยความรู้ ความชำนาญของมนุษย์ เพื่อปฏิบัติงานในขั้นตอนต่าง ๆ นับตั้งแต่การสร้างชุดของศัพท์ควบคุม การวิเคราะห์เนื้อหาเอกสาร และการคัดเลือกคำศัพท์ ส่วนการทำดัชนีจากเนื้อหาเอกสารโดยใช้เครื่องคอมพิวเตอร์นั้นใช้เวลาและค่าใช้จ่ายน้อยกว่า

Harter (1986: 52) ได้สรุปคุณลักษณะของศัพท์ดัชนีทั้ง 2 ประเภทในรูปตารางดังนี้

ตารางที่ 2 แสดงคุณลักษณะของศัพท์ควบคุมและศัพท์อิสระ

ศัพท์อิสระ	ศัพท์ควบคุม
มีความหมายชัดเจน	ไม่มีความหมายชัดเจน
ค้นเรื่องทั่ว ๆ ไปได้ยาก	ค้นเรื่องทั่ว ๆ ไปได้ง่าย
มีจุดเข้าถึงข้อมูลหลายรูปแบบ	มีจุดเข้าถึงข้อมูลได้น้อย
มีปัญหาเกี่ยวกับศัพท์ซึ่งมีความหมายเหมือนกัน	มีการควบคุมศัพท์ซึ่งมีความหมายเหมือนกัน
มีปัญหาเกี่ยวกับคำพ้องรูป	มีการควบคุมคำพ้องรูป

ตารางที่ 2 แสดงคุณลักษณะของศัพท์ควบคุมและศัพท์อิสระ (ต่อ)

ศัพท์อิสระ	ศัพท์ควบคุม
มีปัญหาเกี่ยวกับผลการค้นที่ไม่ตรงกับความต้องการแต่ตรงกับคำค้น	มีการเชื่อมคำค้นในขั้นตอนการกำหนดตรรกะนี้เพื่อแก้ปัญหาผลการค้นที่ไม่ตรงกับความต้องการ
มีความยืดหยุ่นสูง	ไม่มีความยืดหยุ่น
ใช้แทนสภาพความเป็นจริงได้ดี	ใช้แทนสภาพความเป็นจริงได้ไม่ดี
ใช้แทนแง่คิดต่าง ๆ ได้มาก	ใช้แทนแง่คิดต่าง ๆ ได้เพียงอย่างเดียว
ไม่ต้องมีการฝึกอบรมการใช้	ต้องมีการฝึกอบรมการใช้
ใช้แทนแนวคิดใหม่ได้ง่าย	ใช้แทนแนวคิดใหม่ได้ยากหรือไม่ได้เลย
ใช้แทนแนวคิดที่ซับซ้อนได้ง่าย	ใช้แทนแนวคิดที่ซับซ้อนได้ยากหรือไม่ได้เลย
คลุมเครือ ไม่ชัดเจน ไม่มีน้ำหนัก	ไม่คลุมเครือ ชัดเจน แน่นนอน
ไม่มีมาตรฐาน	มีมาตรฐาน
ใช้แสดงความคิดได้อย่างเสรี	ใช้แสดงความคิดได้อย่างจำกัด
ไม่กะทัดรัด	กะทัดรัด
ไม่ต้องมีการทำตรรกะนี้	มีปัญหาเรื่องความไม่คงที่ในการทำตรรกะนี้
ผู้ใช้ต้องคิดคำค้น คำที่มีความหมายเหมือนกันเอง	มีการโยงคำศัพท์ที่ควรใช้เพิ่มเติม
มีความครอบคลุมกว้างขวางและละเอียดสมบูรณ์	ไม่มีความครอบคลุมกว้างขวางและละเอียดสมบูรณ์

(ที่มา : Harter, 1986: 52)

การค้นคืนด้วยศัพท์ควบคุมและศัพท์อิสระ

การค้นคืนข้อมูลเป็นการจําแนกเรื่องที่ใช้ต้องการออกจากเรื่องอื่น ๆ ที่ผู้ใช้ไม่ต้องการ สิ่งที่จะช่วยให้เข้าถึงเรื่องที่ต้องการได้คือ ตรรกะนี้ การค้นคืนโดยใช้ศัพท์ตรรกะนี้เป็นการเปรียบเทียบระหว่างคำศัพท์ที่ผู้ใช้ใช้แทนเรื่องที่ต้องการค้น กับคำศัพท์ที่ผู้สร้างฐานข้อมูลกำหนดขึ้นเพื่อแทนเนื้อหา หรือแนวความคิดของเอกสาร หากพบว่าเหมือนกัน ระบบจะแยกข้อมูลที่ค้นได้ และสามารถแสดงผลการค้นให้ผู้ค้นหา หากข้อมูลที่ค้นได้ตรงกับความต้องการของผู้ใช้ ถือว่าการค้นประสบความสำเร็จ

ความสำเร็จหรือความล้มเหลวในการค้นข้อมูลนั้น ขึ้นอยู่กับปัจจัยต่าง ๆ ดังนี้

(Lancaster, 1979: 176-177)

1. ความสามารถของผู้ทำหน้าที่ค้นคืนในการแปลความต้องการของผู้ใช้
2. ความซับซ้อนของคำถาม

3. ความสามารถของภาษาธรรมชาติ
4. นโยบายการทำธรรมชาติ
5. ประสิทธิภาพของโปรแกรม
6. ความสามารถของผู้ค้นในการคัดเลือกคำศัพท์ และกำหนดกลยุทธ์การค้นคืนที่

เหมาะสม

ปัจจัยที่มักมีการนำไปใช้ในการศึกษาประสิทธิภาพของการค้น คือ ภาษาของธรรมชาติ และเครื่องมือที่นิยมใช้ในการวัดประสิทธิภาพของระบบคือ Recall และ Precision

Lancaster (1989: 18) ได้สำรวจงานวิจัยที่เกี่ยวกับศัพท์ควบคุมและศัพท์อิสระ พบว่า การศึกษาเปรียบเทียบการค้นด้วยศัพท์ควบคุมและศัพท์อิสระในช่วงทศวรรษที่ 1960 ถึง 1970 ปรากฏว่าศัพท์อิสระสามารถหาหน้าที่ได้ดีทัดเทียมกับศัพท์ควบคุม ตั้งแต่ปี 1980 เป็นต้นมา มีการศึกษาในเรื่องทานองเดียวกัน ต่างกันที่ในยุคนี้จะทำการทดลองกับฐานข้อมูลขนาดใหญ่ที่ใช้ปฏิบัติงานจริง

Feng (1989) ได้ศึกษาประสิทธิภาพของภาษาธรรมชาติประเภทศัพท์ควบคุมและศัพท์อิสระ ในการค้นคืนจากฐานข้อมูลในสาขาบรรณารักษศาสตร์และสารนิเทศศาสตร์จำนวน 3 ฐาน คือ Educational Resources Information Center (ERIC) , Library and Information Science Abstracts (LISA) และ Information Science Abstracts (ISA) โดยให้กลุ่มตัวอย่างเลือกคำถาม 1 คำถาม จาก 4 คำถาม ที่กำหนดไว้ล่วงหน้า แล้วนำไปค้นจากฐานข้อมูลทั้ง 3 ฐาน โดยแต่ละฐานจะต้องค้น 5 รูปแบบดังนี้

1. ค้นด้วยศัพท์อิสระ และศัพท์อิสระร่วมกับศัพท์ควบคุม โดยค้นจากชื่อเรื่อง
2. ค้นด้วยศัพท์อิสระ และศัพท์อิสระร่วมกับศัพท์ควบคุม โดยค้นจากสาระสังเขป
3. ค้นด้วยศัพท์อิสระ และศัพท์อิสระร่วมกับศัพท์ควบคุม โดยค้นจากชื่อเรื่องและสาระสังเขป
4. ค้นด้วยศัพท์ควบคุม โดยค้นจากเขตข้อมูลศัพท์ควบคุม
5. ค้นด้วยศัพท์อิสระ และศัพท์อิสระร่วมกับศัพท์ควบคุม โดยค้นทั้งระเบียบ

เครื่องมือที่ใช้วัดผลการค้นคือ Recall และ Precision พร้อมวัดระดับความตรงของข้อมูลที่ค้นได้ 3 ระดับ คือ R (Relevance) คือข้อมูลที่เกี่ยวข้อง P (Partially relevance) คือข้อมูลที่เกี่ยวข้องเป็นบางส่วน และ I (Irrelevance) คือข้อมูลที่ไม่ตรง ผลจากการศึกษาพบว่า การค้นด้วยศัพท์อิสระได้ผลจำนวนจำกัด ส่วนศัพท์ควบคุมได้ผลการค้นพอประมาณ การค้นที่ใช้ศัพท์ควบคุมและศัพท์อิสระพร้อม ๆ กัน ผลการค้นจะดีกว่าการค้นด้วยศัพท์ควบคุมหรือศัพท์อิสระ

Peter และ Kurth (1991) ได้ศึกษาการใช้ศัพท์ควบคุม และศัพท์อิสระจากชื่อเรื่องของผู้ใช้บริการบรรณารายการแบบออนไลน์ของมหาวิทยาลัยมิสซูรี เมืองแคนซัส ประเทศ

สหรัฐอเมริกา โดยศึกษาจากข้อมูลซึ่งเกิดจากการประมวลผลด้วยเครื่องคอมพิวเตอร์ (Transaction logs) ตั้งแต่ เดือนกุมภาพันธ์ถึงมีนาคม 1990 ซึ่งมีข้อมูลเกี่ยวกับรายการทางบรรณานุกรมที่ได้รับการค้นคืน ผลจากการศึกษาปรากฏว่าการค้นจำนวนร้อยละ 6.6 ผู้ใช้บริการใช้ทั้งศัพท์ควบคุมและศัพท์อิสระ และในจำนวนนั้น มีการค้นร้อยละ 58 ที่เริ่มการค้นด้วยศัพท์อิสระ โดยทั่วไปแล้วผู้ใช้อย่างคงพอใจกับการค้นด้วยศัพท์ควบคุม ในขณะที่ผลการค้นด้วยศัพท์อิสระมีจำนวนมากกว่า

Rowley (1990) ได้ศึกษาการค้นด้วยภาษาธรรมชาติและภาษาควบคุมโดยให้กลุ่มตัวอย่างจำนวน 40 คน สร้างฐานข้อมูลขนาดเล็ก แต่ละระเบียบประกอบด้วยเขตข้อมูล 3 เขต คือชื่อเรื่อง หัวเรื่องจาก Sears' List of Subject Headings และศัพท์จาก ERIC Thesaurus การทดลองค้นข้อมูลของกลุ่มตัวอย่างมี 2 รูปแบบ คือ ค้นด้วยศัพท์ 1 คำ และ คำศัพท์ 2 คำ โดยมี "and" เชื่อม การค้นมี 3 ลักษณะ คือ ค้นโดย 1) ใช้ศัพท์อิสระจากชื่อเรื่อง 2) ใช้ศัพท์ควบคุมจาก Sears' List of Subject Headings และ 3) ใช้หัวเรื่องจาก ERIC Thesaurus ผลที่ได้จากการทดลองจะนำไปวิเคราะห์หาค่า Recall และ Precision ซึ่งผลจากการวิเคราะห์ปรากฏว่า การค้นด้วยศัพท์อิสระ (คือจากชื่อเรื่อง) ค่าของ Recall จะต่ำกว่าการค้นด้วยศัพท์ควบคุม (คือศัพท์จาก Sears และ ERIC) แต่หากพิจารณา ค่าของ Precision แล้ว พบว่า การค้นด้วยศัพท์อิสระ ค่าของ Precision จะสูงกว่าการค้นด้วยศัพท์ควบคุม

พัชรี พันดาวงษ์ (2536) ได้สร้างระบบการจัดเก็บและค้นคืนวัสดุจดหมายเหตุของโครงการจดหมายเหตุ มหาวิทยาลัยศิลปากร โดยใช้โปรแกรมสำเร็จรูป Mini-Micro CDS/ISIS ซึ่งสามารถค้นคืนข้อมูลได้ด้วยตรรกะนิเสธประเภทศัพท์ควบคุม คือหัวเรื่อง และตรรกะนิเสธประเภทศัพท์อิสระจากคำในชื่อเรื่อง สารสังเขป และหมายเหตุ ผลการประเมินการค้นคืนโดยการคำนวณหาค่า Recall และ Precision ปรากฏว่า ถ้าค่า Recall เพิ่มขึ้น ค่า Precision จะเพิ่มขึ้นด้วย

นอกจากภาษาตรรกะนี้แล้ว ปัจจัยอีกประการหนึ่งที่มีผลต่อความสำเร็จหรือความล้มเหลวในการค้นคืน คือ ความสามารถของผู้ค้นในการคัดเลือกคำศัพท์ ในการค้นคืนข้อมูลนั้นคำศัพท์ที่ผู้ใช้เลือกใช้จะต้องตรงกับคำศัพท์ตรรกะของฐานข้อมูล และต้องเป็นคำที่หมายถึงเรื่องเดียวกัน อย่างไรก็ตาม เรื่องเดียวกันอาจมีคำศัพท์ต่างกัน ดังนั้นในการค้นข้อมูลเรื่องเดียวกันโดยผู้ค้นคนละคน ผลการค้นจะแตกต่างกันเพราะใช้คำค้นต่างกัน (Saracevic and Kanter 1988; Pao, 1989)

ดังนั้นการเลือกใช้คำค้นของผู้ใช้นั้น จึงเป็นเรื่องที่มีความสำคัญ ได้มีผู้ศึกษาเกี่ยวกับการใช้คำศัพท์ของผู้ค้น ว่าตรงกับศัพท์ตรรกะในฐานข้อมูลหรือไม่ ตลอดจนมีการหาวิธีปรับปรุงตรรกะนิเสธประเภทศัพท์ควบคุม เพื่อให้ตรงกับศัพท์ที่ผู้คิดค้นเองมากขึ้น

ในวิทยานิพนธ์ของ Lester (Lester, 1989 quoted in Connell, 1991) ซึ่งศึกษาเกี่ยวกับการปรับปรุงการเข้าถึงฐานข้อมูลออนไลน์โดยใช้เรื่องเป็นคำค้น Lester ได้ศึกษาข้อมูลการประมวลผลด้วยคอมพิวเตอร์ (Transaction logs) แล้วทำการเปรียบเทียบคำค้นของผู้ใช้กับหัวเรื่องจากคู่มือ Library of Congress Subject Headings (LCSH) ผลปรากฏว่าคำค้นและหัวเรื่องดังกล่าวตรงกันร้อยละ 40

Connell (1991) ได้ศึกษาการค้นคืนในระบบออนไลน์ที่ใช้เรื่องเป็นคำค้น โดยการรวบรวมเรื่องย่อของหนังสือจาก Book Review Digest จำนวน 1,023 รายการ เพื่อศึกษาว่าคำศัพท์ในเรื่องย่อ ตรงกับหัวเรื่องใน LCSH หรือคำในชื่อเรื่องเพียงใด ผลปรากฏว่า คำในเรื่องย่อตรงกับหัวเรื่องหลักร้อยละ 35.7 และตรงกับคำในชื่อเรื่องร้อยละ 27.8 แต่เมื่อใช้คำศัพท์ทั้ง 3 ประเภท พร้อม ๆ กัน ทำให้ค่าของ Recall สูงขึ้นถึงร้อยละ 20 คือสูงขึ้นจากเดิมร้อยละ 67 เป็น 87

สำหรับการค้นข้อมูลบัตรรายการในระบบออนไลน์ด้วยศัพท์อิสระนั้น Ensor (1992) ได้ศึกษาการใช้คำค้นจากชื่อเรื่องของผู้ใช้ระบบห้องสมุดอัตโนมัติ ที่มีชื่อว่า NOTIS (Northwestern Online Total Integrated Systems) ในห้องสมุดของ Indiana State University ประเทศสหรัฐอเมริกา ผลการวิจัยพบว่า กว่าครึ่งหนึ่งจำนวนการค้นผู้ค้นใช้ศัพท์จากชื่อเรื่อง และเห็นว่าการใช้ศัพท์จากชื่อเรื่องนั้นเรียนรู้ได้ง่าย

จะเห็นได้ว่าภาษาธรรมชาติและการเลือกใช้คำค้น มีผลต่อความสำเร็จหรือความล้มเหลวโดยตรง มีความเชื่อกันว่า ในกระบวนการค้นคืน ข้อมูลอย่างน้อยร้อยละ 35 ได้หายไป ตั้งแต่ผู้ใช้บริการติดต่อขอข้อมูล จนกระทั่งได้รับข้อมูล โดยที่ร้อยละ 10 ของข้อมูลหายไปเนื่องจากผู้ใช้ไม่สามารถอธิบายความต้องการได้อย่างถูกต้อง ส่วนอีกร้อยละ 25 ได้หายไปเนื่องจากการที่ไม่สามารถใช้คำค้นได้เฉพาะเจาะจงและถูกต้อง รวมทั้งการทาดรรชนี ที่เอกสารไม่ได้รับการทาดรรชนีอย่างถูกต้อง (ชลีย์พร เหมะรัชตะ, 2535: 22)

ดรรชนีจึงมีบทบาทสำคัญในการจัดเก็บและการค้นคืนสารสนเทศ ประสิทธิภาพของระบบการค้นคืนข้อมูลจึงขึ้นอยู่กับคุณภาพของการจัดทาดรรชนี