

บทที่ 4

การคัดแยกข้อความจากโครงสร้างต้นไม้มองไวยากรณ์

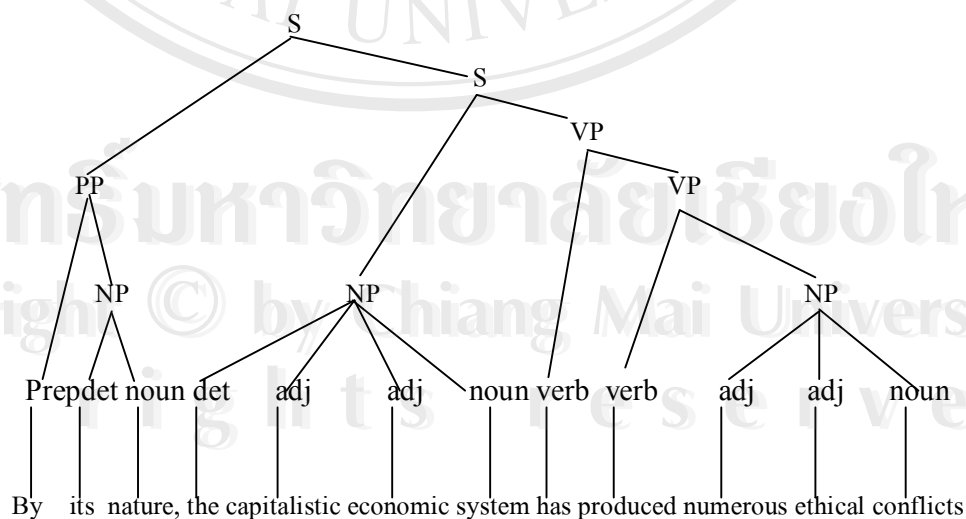
สำหรับจุดมุ่งหมายของงานวิจัยนี้มีความต้องการในการคัดแยกข้อความที่สนใจจากบทความภาษาอังกฤษ ซึ่งการทำงานกับภาษาธรรมชาติก็นำมาซึ่งปัญหาเกี่ยวกับการวิเคราะห์โครงสร้างของภาษา โดยโครงสร้างของประโยคที่มีความกำกวมจะมีผลต่อความถูกต้องของการคัดแยกข้อความที่ต้องการ การคัดแยกข้อความจากบทความที่มีความกำกวมโดยตรงนั้นจะทำให้ได้ข้อความที่ไม่ถูกต้องอันเป็นผลเนื่องมาจากความกำกวมนั้น ซึ่งการคัดแยกข้อความแบบเดิมจะใช้วิธีการทำแพทเทิร์นแมชชีนซึ่งเป็นการค้นหาข้อความที่สนใจจากคำสำคัญที่พบและถือว่าข้อความทั้งหมดที่อยู่หน้าและหลังคำสำคัญนั้นเป็นประธานและกรรมของประโยค ข้อความที่ได้จากวิธีการข้างต้นนั้นอาจทำให้ได้ข้อความที่เป็นส่วนของคำขยายที่ไม่จำเป็นติดมาด้วย งานวิจัยนี้จึงได้นำเสนอการใช้พาสเซอร์ในการวิเคราะห์โครงสร้างของประโยคก่อนที่จะทำการคัดแยกข้อความที่ต้องการ เพื่อให้ได้ข้อความที่กระชับแต่มีความหมายมากขึ้น เนื่องจากพาสเซอร์นั้นสามารถทำการวิเคราะห์ประโยคและจัดโครงสร้างของประโยคที่ทำการวิเคราะห์ได้ให้มีรูปแบบเป็นโครงสร้างแบบต้นไม้ ทำให้สามารถคัดแยกข้อความที่ต้องการได้โดยอาศัยโครงสร้างของประโยคที่วิเคราะห์ได้จากพาสเซอร์ ซึ่งผลลัพธ์ที่ได้จากการคัดแยกโดยการวิเคราะห์โครงสร้างของประโยคนั้นจะทำให้สามารถแยกส่วนขยายของประธานและกรรมออกจากประธานและกรรมของประโยคดังกล่าวได้

สำหรับลักษณะของบทความที่นำมาใช้ในการทดสอบในครั้งนี้เป็นบทคัดย่อของงานวิจัยในสาขาสังคมวิทยาที่เกี่ยวข้องกับจริยธรรมทางธุรกิจ เปรียบเทียบกับบทความที่เป็นบทคัดย่อของงานวิจัยในสาขาไบโอเคมี ซึ่งหลักการในการเขียนนั้นจะต้องมีความถูกต้องทางไวยากรณ์ สำหรับบทคัดย่อทั้งสองชุดนั้นเป็นบทความทางวิชาการดังนั้นการเขียนจึงมีคำศัพท์เฉพาะปรากฏอยู่ด้วย โดยเฉพาะบทคัดย่อในสาขาไบโอเคมี โดยประโยคส่วนใหญ่ในบทความนั้นจะเป็นประโยคความรวม ประโยคความซ้อน ซึ่งจะมีประโยคย่อยเป็นส่วนขยายทั้งภาคประธานและภาคแสดง คำกริยาที่ใช้มักจะเป็นคำกริยาผสม มีลักษณะเป็นประโยคซ้อนประโยค (compound – complex sentence) ประโยคจึงมีความยาวมาก ค่อนข้างซับซ้อน และมักจะมีการใช้คำเชื่อมอยู่มาก บางประโยคอาจมีประธานและกริยามากกว่าหนึ่งก็ได้ เช่น “The ethics of rights or the separative model has dominated Western thought since the Enlightenment and the ethics of care was developed as a feminist critique seeking to rebalance our basic thought structure” หรือ “The transformation of

companies into post-bureaucratic forms of organization such as networks raises new implementation challenges for ethics in business” เป็นต้น

ถ้าเรานูमानจากโครงสร้างของประโยคว่าข้อความส่วนที่อยู่ด้านหน้าของกริยาหลักนั้น เป็นภาคประธานและส่วนที่อยู่ด้านหลังของกริยาหลักนั้นจะทำหน้าที่เป็นกรรมของประโยค จาก ลักษณะประโยคในบทคัดย่อที่กล่าวมาแล้วนั้น พบว่าประโยคส่วนใหญ่จะมีนามวลีในลักษณะที่เป็นอนุประโยคยาว ๆ ทำหน้าที่เป็นประธานและกรรมของประโยค เช่นประโยค “The transformation of companies into post-bureaucratic forms of organization, such as networks, raises new implementation challenges for ethics in business” จากประโยคดังกล่าวเมื่อคำกริยาคือ raises พบว่าประธานของประโยคคือ The transformation of companies into post-bureaucratic forms of organization, such as networks ซึ่งมีลักษณะเป็นกลุ่มคำนามที่ทำหน้าที่เป็นประธานของ ประโยค หรือ ประโยค “This research contributes a set of generic models that examine business ethics dynamics” เมื่อคำกริยาคือ contributes พบว่ากรรมของประโยค คือ a set of generic models that examine business ethics dynamics ซึ่งมีลักษณะเป็นอนุประโยค

การแบ่งวลีในประโยคนั้นจะแบ่งตามหน้าที่ของวลีที่มีในประโยค ซึ่งอาจจะเป็น นามวลี กริยาวลี หรือ บุพบทวลี โครงสร้างของวลีจะช่วยให้ในการแสดงการจัดกลุ่มของกลุ่มคำในประโยค โดยแสดงเป็นลำดับชั้นขององค์ประกอบต่าง ๆ ซึ่งอาจแสดงได้หลายรูปแบบแตกต่างกันออกไป รวมทั้งสามารถแสดงให้อยู่ในรูปของโครงสร้างต้นไม้ได้ด้วย ดังตัวอย่างเช่น ประโยค “By its nature, the capitalistic economic system has produced numerous ethical conflicts” สามารถแสดง เป็นโครงสร้างต้นไม้ได้ดังรูป 4.1



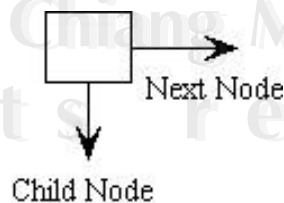
รูป 4.1 แสดงโครงสร้างของประโยค

จากโครงสร้างต้นไม้ดังกล่าวทำให้เกิดแนวความคิดในการวิเคราะห์โครงสร้างของประโยค และจัดโครงสร้างของประโยคดังกล่าวให้อยู่ในรูปของโครงสร้างต้นไม้ก่อนที่จะทำการคัดแยกข้อความเพื่อลดความกำกวมของประโยคและเพื่อให้สามารถท่องลงไปโครงสร้างต้นไม้และทำการคัดแยกข้อความที่ต้องการได้

สำหรับพาสเชอร์นั้นเป็นส่วนประกอบหนึ่งของลิงค์แกรมมาพาสเชอร์ ลิงค์แกรมมาพาสเชอร์เป็น “grammatical parsing” มีพื้นฐานการทำงานอยู่บนหลักการวากยสัมพันธ์ (syntactic parser) ซึ่งเป็นความสัมพันธ์กันระหว่างคำในประโยคตามหลักการของภาษาอังกฤษ โดยพาสเชอร์จะนำความสัมพันธ์ของลิงค์ที่ได้จากลิงค์แกรมมาไปใช้ในการวิเคราะห์โครงสร้างของประโยค ดังนั้นเราจึงสามารถจัดกลุ่มของวลีและประโยคย่อยที่มีในประโยคได้ เพื่อให้สามารถทำการคัดแยกข้อความที่ต้องการโดยใช้ประโยชน์จากโครงสร้างของประโยคที่วิเคราะห์ได้ ซึ่งมีวิธีการดำเนินการดังที่จะกล่าวต่อไปนี้

4.1 โครงสร้างต้นไม้ไวยากรณ์

จากการทำงานของลิงค์แกรมมาจะทำให้ได้ลิงค์ที่ทำหน้าที่แสดงความสัมพันธ์กันของคำแต่ละคู่ในประโยคนั้น ๆ พาสเชอร์จึงนำลิงค์ที่ได้จากลิงค์แกรมมานั้นมาใช้ในการวิเคราะห์องค์ประกอบต่าง ๆ ของประโยคซึ่งเป็นไปตามกระบวนการที่ได้อธิบายไปแล้วในบทที่ 3 ซึ่งพาสเชอร์จะทำการวิเคราะห์โครงสร้างของประโยคออกเป็นกลุ่มวลีต่าง ๆ ที่ประกอบกันขึ้นเป็นประโยคนั้น องค์ประกอบส่วนต่าง ๆ ของประโยคไม่ว่าจะเป็นนามวลี กริยาวลี และบุพบทวลีนั้นถือว่าเป็นส่วนสำคัญของประโยค โดยเชื่อว่าแต่ละวลีประกอบไปด้วยกลุ่มคำที่มีความหมายและมีหน้าที่ในประโยค ซึ่งวลีดังกล่าวอาจทำหน้าที่เป็นประธาน กรรม หรือส่วนขยายของประโยค โดยโครงสร้างของประโยคที่ได้จากการวิเคราะห์โดยพาสเชอร์นั้นจะถูกจัดให้อยู่ในรูปของโครงสร้างแบบต้นไม้ ซึ่งโครงสร้างต้นไม้ดังกล่าวจะประกอบไปด้วยโหนดแต่ละโหนด ซึ่งในที่นี้เรียกว่า “CNode” ซึ่งประกอบไปด้วยตัวชี้ 2 ตัวด้วยกันคือ ตัวชี้ที่ชี้ไปยังโหนดถัดไป และตัวชี้ที่ชี้ไปยังโหนดลูก แสดงดังรูปที่ 4.2



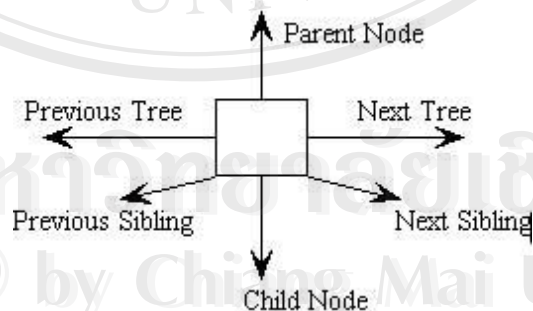
รูปที่ 4.2 แสดงโครงสร้าง Cnode

จากนั้นนำโครงสร้างของประโยคที่วิเคราะห์ได้และอยู่ในโครงสร้างต้นไม้ดังกล่าวมาจัดให้อยู่ในรูปของสายอักขระที่แบ่งกลุ่มวลีด้วยเครื่องหมายวงเล็บ พร้อมทั้งมีตัวอักษรแสดงวลีต่าง ๆ กำกับด้วย เช่น “NP” หมายถึง นามวลี “VP” หมายถึง กริยวลี และ “PP” หมายถึง บุพบทวลี

ยกตัวอย่าง ประโยคนำเข้าเป็น “By its nature, the capitalistic economic system has produced numerous ethical conflicts” โครงสร้างของประโยคที่วิเคราะห์ได้จะถูกจัดให้อยู่ในรูปแบบดังนี้

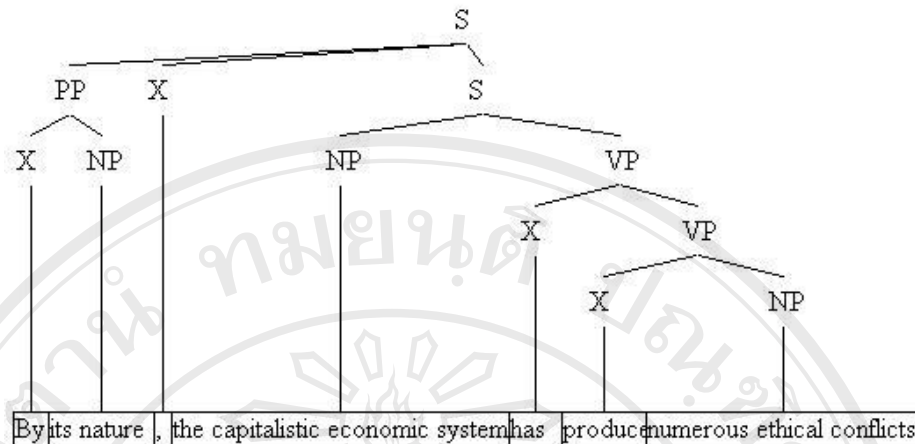
“S (PP (X (By) NP (its nature)) X (,) S (NP (the capitalistic economic system) VP (X (has) VP (X (produced) NP (numerous ethical conflicts))))))”

ซึ่งจากผลลัพธ์ที่ได้แต่ละวลีจะถูกจัดกลุ่มไว้ด้วยเครื่องหมายวงเล็บและกำกับไว้ด้วยตัวอักษรที่แสดงวลีในวงเล็บนั้น เช่น “S” หมายถึง ประโยค หรือ อนุประโยค “NP” หมายถึง นามวลี “PP” หมายถึง บุพบทวลี เป็นต้น ผลลัพธ์ที่จัดในรูปของอักขระนี้จะนำไปเป็นข้อมูลเพื่อจัดการให้โครงสร้างของประโยคดังกล่าวนั้นอยู่ในรูปของโครงสร้างต้นไม้ที่มีการเปลี่ยนแปลงโครงสร้างของแต่ละโหนดใหม่ เนื่องจากโครงสร้างของโหนดเดิมที่เป็น “CNode” นั้นประกอบไปด้วยตัวชี้เพียง 2 ตัวเท่านั้นทำให้ไม่สามารถย้อนกลับไปยังโหนดก่อนหน้านี้ภายในโครงสร้างต้นไม้ได้ ดังนั้นในงานวิจัยนี้จึงได้ทำการเปลี่ยนแปลงโครงสร้างของโหนดใหม่ซึ่งพัฒนาโดยใช้ภาษาซี พลัส พลัส บิวเดอร์ (C++ Builder) ซึ่งโครงสร้างใหม่มีลักษณะดังนี้ ให้แต่ละโหนดในโครงสร้างต้นไม้มีตัวชี้ทั้งหมด 6 ตัวทำหน้าที่ชี้ไปยังโหนดต่าง ๆ ในโครงสร้างต้นไม้ ประกอบด้วยตัวชี้ที่ชี้ไปยังต้นไม้ก่อนหน้า ตัวชี้ที่ชี้ไปยังต้นไม้ถัดไป ตัวชี้ที่ชี้ไปยังโหนดถัดไป ตัวชี้ที่ชี้ไปยังโหนดก่อนหน้า ตัวชี้ที่ชี้ไปยังโหนดพ่อแม่และตัวชี้ที่ชี้ไปยังโหนดลูก ดังแสดงดังรูป 4.3



รูปที่ 4.3 แสดงโครงสร้าง Tree Node

จากประโยคตัวอย่างข้างต้นโครงสร้างของประโยคที่ได้ถูกจัดให้อยู่ในโครงสร้างต้นไม้ดังรูปที่ 4.4



รูปที่ 4.4 แสดงโครงสร้างต้นไม้ไวยากรณ์

โดยที่มีรากโหนด (root node) เป็นโหนดเริ่มต้นที่กำกับด้วยตัวอักษร S สำหรับรากโหนดนั้นตัวชี้ที่ชี้ไปยังโหนดพ่อแม่ซึ่งจะมีค่าเป็น “null” เสมอ สำหรับโหนดใบ (leaves node) จะเก็บกลุ่มวลีต่าง ๆ ไว้โดยมีโหนดพ่อแม่ทำหน้าที่เก็บตัวอักษรที่ระบุชนิดของวลีไว้ สำหรับโหนดที่มีตัวอักษร “X” กำกับนั้นเป็นการแสดงส่วนหลักของวลี โดยอาศัยทฤษฎี “X-bar” ซึ่งมีประโยชน์ในการระบุส่วนหลักและส่วนขยายของวลี เนื่องจากทฤษฎี “PSG: Phrase Structure Grammar” นั้นจะแบ่งโครงสร้างของวลีได้เพียงแค่ 2 ระดับคือระดับคำและระดับวลี ซึ่งไม่ครอบคลุมในส่วนที่เป็นอักษรแต่ไม่ได้เป็นวลีเต็ม ดังนั้นจึงได้มีการนำเสนอทฤษฎี X-bar เพื่อแสดงส่วนที่เป็นอักษรซึ่งมีระดับอยู่ระหว่างระดับคำและระดับวลี เพื่อช่วยในการแสดงหน้าที่ที่แตกต่างกันในวลีเดียวกันเมื่อประโยคมีความซับซ้อนมากขึ้น

4.2 การคัดแยกข้อความจากโครงสร้างต้นไม้

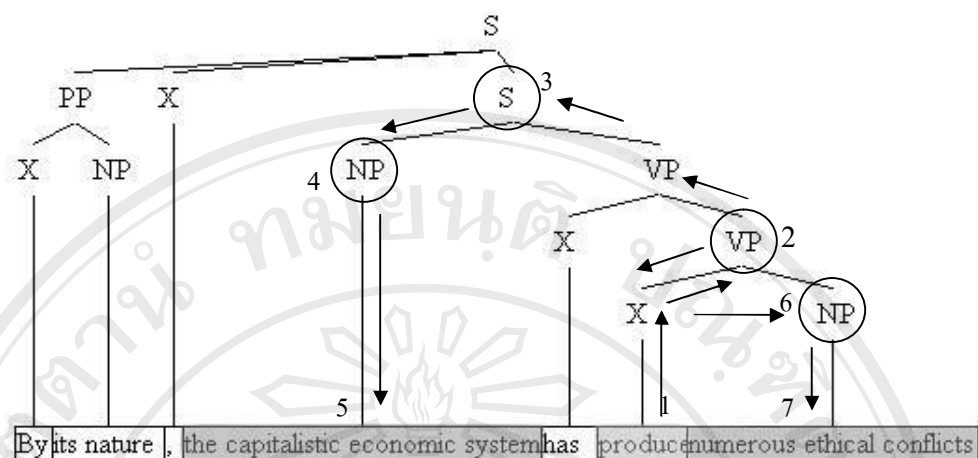
เมื่อโครงสร้างของประโยคดังกล่าวถูกจัดให้อยู่ในรูปของโครงสร้างต้นไม้แล้ว ในการท่องลงไปในโครงสร้างต้นไม้เพื่อทำการคัดแยกข้อความที่ต้องการนั้น อันดับแรกจะต้องมีการกำหนดคอนเซ็ปต์แพตเทิร์นซึ่งได้อธิบายในบทที่ 2 โดยคอนเซ็ปต์แพตเทิร์นประกอบไปด้วยคำสำคัญและรูปแบบทางภาษา (linguistic pattern) เช่น กำหนดว่าต้องมีรูปแบบเป็นนามวลี เป็นต้น คำสำคัญที่ปรากฏอยู่ในคอนเซ็ปต์แพตเทิร์นนั้นจะเป็นทริกเกอร์เวิร์ดที่ใช้ในการค้นหาข้อความที่ต้องการในประโยค ซึ่งโดยทั่วไปทริกเกอร์เวิร์ดจะเป็นคำกริยาและข้อความที่ต้องการคัดแยกนั้นจะเป็นนามวลีที่ทำหน้าที่เป็นประธานหรือกรรมในประโยค สำหรับการระบุคำสำคัญนั้นจะได้มาจากข้อมูลที่ต้องการคัดแยกจากบทความ โดยจะทำการพิจารณาคำกริยาที่เกี่ยวข้องกับข้อความที่เป็นจุดมุ่งหมายในการคัดแยก ซึ่งในการวิจัยในครั้งนี้ได้ทำการทดลองกับบทคัดย่อของงานวิจัยทั้งงานวิจัยในสาขา

สังคมวิทยาและงานวิจัยในสาขาไบโอเคมี ดังนั้นในงานวิจัยนี้จึงมีจุดมุ่งหมายในการคัดแยกข้อความที่เกี่ยวข้องกับงานวิจัยในแต่ละบทคัดย่อนั้น เช่น ต้องการทราบว่าบทคัดย่อนั้นทำงานวิจัยเรื่องใด ทำอย่างไร ได้ผลจากการวิจัยนั้นอย่างไร ดังนั้นคำสำคัญที่เกี่ยวข้องกับข้อมูลที่ต้องการคัดแยกจากบทคัดย่อที่ใช้ในงานวิจัยครั้งนี้จึงเป็นคำกริยาที่ปรากฏในคอนเซ็ปต์แพดเทิร์น ยกตัวอย่างเช่น analyzes, addressed, present, found, produced, support, used, attacked เป็นต้น โดยข้อมูลที่คัดแยกได้นั้นจะต้องเป็นนามวลีที่มีความสัมพันธ์กับคำสำคัญที่กำหนดในคอนเซ็ปต์แพดเทิร์น ซึ่งนามวลีที่คัดแยกได้ดังกล่าวจะมีความสำคัญโดยอาจจะเป็นประธานหรือกรรมในประโยคนั้น นอกจากนี้คอนเซ็ปต์แพดเทิร์นยังใช้ในการตรวจสอบความถูกต้องของข้อมูลที่ทำกรคัดแยกมาได้ว่าตรงกับคอนเซ็ปต์แพดเทิร์นที่กำหนดหรือไม่ เช่น กำหนดว่าเป็นนามวลีที่อยู่หน้าคำสำคัญ เพื่อทำหน้าที่เป็นประธานของประโยค เป็นต้น

ในการสืบค้นนั้นคำสำคัญที่ถูกกำหนดในคอนเซ็ปต์แพดเทิร์นจะถูกค้นหาในแต่ละประโยค และจะทำการทอกลงไปเพื่อค้นหานามวลีที่มีความเกี่ยวข้องกับคำสำคัญนั้นเฉพาะในโครงสร้างต้นไม้ของประโยคที่มีคำสำคัญปรากฏอยู่เท่านั้น เมื่อพบคำสำคัญปรากฏอยู่ในโครงสร้างของประโยคใดให้ทำการตรวจสอบว่าคำสำคัญนั้นมีไหนดพ่อนแม่เป็นกริยาวลีหรือไม่ เพื่อตรวจสอบตามเงื่อนไขที่กำหนดว่าคำสำคัญนั้นจะต้องเป็นคำกริยา เมื่อทำการตรวจสอบเรียบร้อยแล้วจะทำการสืบค้นไปสองทางด้วยกัน คือสืบค้นหานามวลีด้านหน้าของทริกเกอร์เวิร์ด โดยสืบค้นย้อนขึ้นไปยังไหนดพ่อนแม่จนกระทั่งเจอไหนด “S” แล้วจึงเริ่มทำการค้นหานามวลีที่ถูกกำกับด้วยไหนด “NP” ต่อไปโดยทำการค้นหาแบบขวาง (breadth-first search) สำหรับการค้นหานามวลีที่ตามหลังคำสำคัญนั้นจะเริ่มค้นที่ “VP” ซึ่งเป็นไหนดพ่อนแม่ของคำสำคัญที่พบ จากนั้นจึงสืบค้นลงไปแล้วไปยังไหนดทางด้านขวาเพื่อค้นหานามวลีที่ถูกกำกับด้วยไหนด “NP” หรือไหนด “PP” ในกรณีที่ต้องการหาบุพบทวลี โดยทำการค้นหาแบบขวางเช่นกัน และเมื่อพบไหนดของนามวลีที่กำกับด้วยตัวอักษร “NP” หรือไหนดบุพบทวลีที่กำกับด้วยอักษร “PP” ให้ทำการดึงข้อความที่อยู่ในไหนดลูก ซึ่งเป็นข้อความที่ต้องการมาแสดงโดยทำการค้นหาแบบลึก (depth-first search) ซึ่งจากประโยคตัวอย่าง “By its nature, the capitalistic economic system has produced numerous ethical conflicts” มีคอนเซ็ปต์แพดเทิร์น ดังนี้

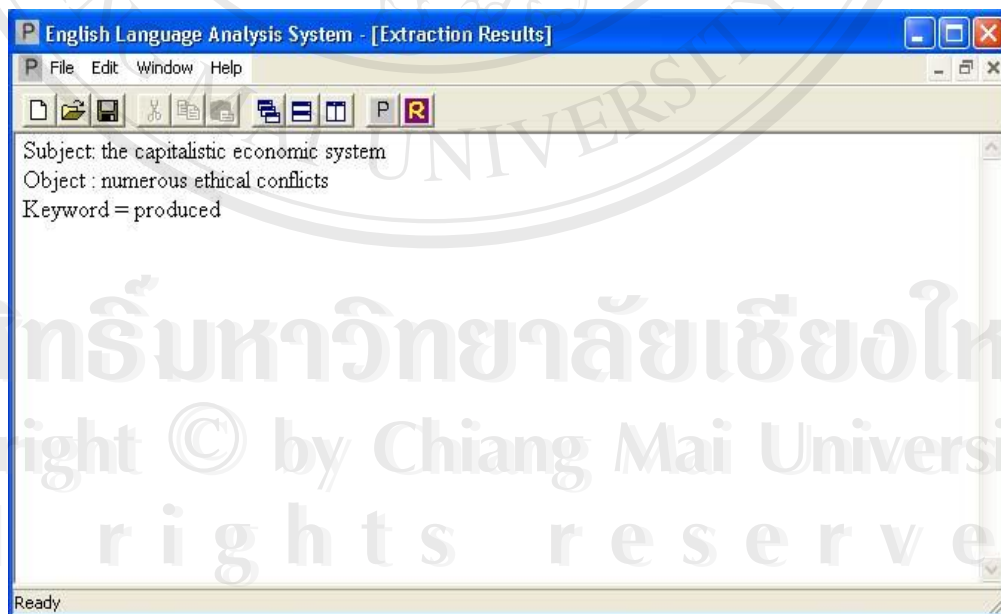
<NP> has produced <NP>

จากคอนเซ็ปต์แพดเทิร์นดังกล่าวเป็นการกำหนดว่า คำสำคัญคือ produced และข้อความที่คัดแยกได้จะต้องเป็นนามวลีที่อยู่หน้าและหลังคำสำคัญ การทอกลงไปในโครงสร้างต้นไม้เพื่อค้นหาคำสำคัญและคัดแยกข้อความที่ต้องการนั้นแสดงดังรูป 4.5 โดยตัวเลขที่กำกับนั้นแสดงลำดับในการทอกลงไปในโครงสร้างต้นไม้



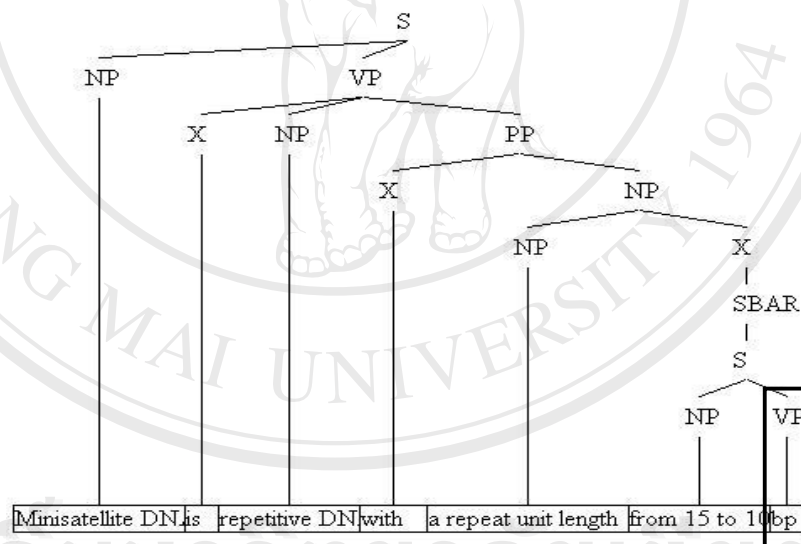
รูปที่ 4.5 แสดงการทอกลงไปในโครงสร้างต้นไม้ของประโยค

ผลที่ได้จากการคัดแยกข้อความจากประโยคนี้นี้จะได้นามวลีที่ต้องการคือ “the capitalistic economic system” ซึ่งเป็นนามวลีที่อยู่หน้าทริกเกอร์เวิร์ดและอนุมานได้ว่าคำนามที่อยู่ด้านหน้าของคำกริยาหลักนั้นเป็นประธานของประโยคจึงนำผลที่ได้จากการคัดแยกนี้เติมลงในช่องของประธาน และ “numerous ethical conflicts” มีตำแหน่งอยู่ด้านหลังของกริยาหลักจึงอนุมานว่าเป็นกรรมของประโยคดังนั้นจึงนำผลที่คัดแยกได้นี้เติมลงในช่องของกรรม ดังแสดงในรูปที่ 4.6



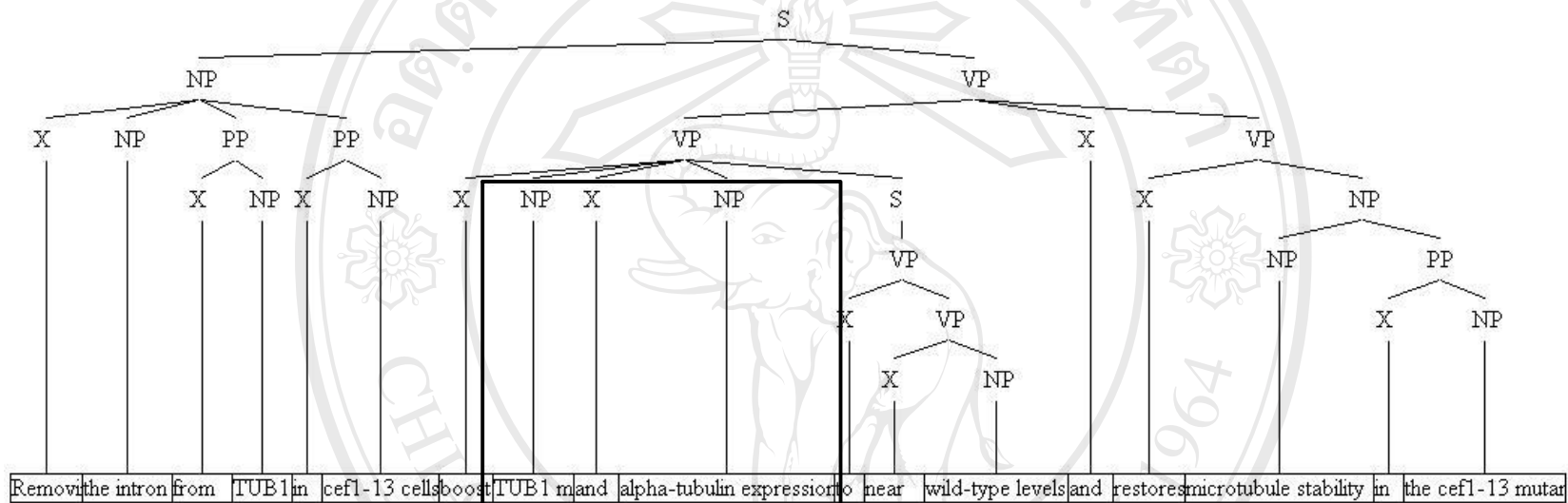
รูป 4.6 แสดงผลที่ได้จากการคัดแยก

รูปแบบของประโยคที่มีผลทำให้ระบบคิดแยกนี้ไม่ได้ประสิทธิภาพ มักจะเป็นประโยคที่ประกอบไปด้วยอนุประโยคยาว ๆ ที่ทำหน้าที่เป็นส่วนขยายของประโยค ประโยคที่มีการใช้คำศัพท์ที่เป็นคำเฉพาะจำนวนมาก มีการใช้สัญลักษณ์ทางคณิตศาสตร์ และสูตรเคมีต่าง ๆ ที่เป็นสัญลักษณ์และตัวเลข และบางประโยคที่มีการใช้คำเชื่อมในประโยค เช่น “and” “or” หรือ “but” หรือการใช้เครื่องหมายจุดภาค อัมภาค ทวิภาค ในประโยค รูปแบบของประโยคดังกล่าวจะมีผลให้พาสเชอร์ทำการวิเคราะห์โครงสร้างของประโยคผิดพลาดได้ ซึ่งเมื่อพาสเชอร์นั้นทำการวิเคราะห์โครงสร้างของประโยคผิดพลาดก็จะส่งผลให้การคัดแยกข้อมูลที่ต้องการนั้นไม่ถูกต้องเช่นกัน ตัวอย่างของประโยคที่มีการวิเคราะห์ผิดพลาด เช่น “Minisatellite DNA is repetitive DNA with a repeat unit length from 15 to 100 bp” เมื่อพาสเชอร์ทำการวิเคราะห์ประโยคจะได้ผลลัพธ์ดังนี้ “S(NP(Minisatellite DNA) VP(X(is) NP(repetitive DNA) PP(X(with) NP(NP(a repeat unit length) X(SBAR(S(NP(from 15 to 100) VP(bp)))))))))” จากผลลัพธ์ที่ได้เห็นว่าพาสเชอร์ทำการวิเคราะห์ว่า “bp” เป็นกิริยาวิลี ซึ่งที่ถูกต้อง “bp” เป็นหน่วยวัดหน่วยหนึ่ง แสดงดังรูป 4.7



รูป 4.7 แสดงโครงสร้างของประโยคที่วิเคราะห์ผิดพลาด

หรือประโยค “Removing the intron from TUB1 in cef1-13 cells boosts TUB1 mRNA and alpha-tubulin expression to near wild-type levels and restores microtubule stability in the cef1-13 mutant” ซึ่งเป็นประโยคที่มีความซับซ้อน เนื่องจากเป็นประโยคที่มีการใช้คำเชื่อม “and” อยู่หลายแห่ง ทำให้เกิดการวิเคราะห์โครงสร้างที่มีผลให้ทำการคัดแยกผิดพลาดได้แสดงดังรูป 4.8 ซึ่งจะแสดงผลที่ได้จากการคัดแยกในหัวข้อ 4.3 ต่อไป



รูป 4.8 แสดงโครงสร้างของประโยคที่มีผลให้ทำการตัดแยกข้อความนั้นผิดพลาด

4.3 แสดงตัวอย่างการคัดแยกข้อความที่ต้องการจากโครงสร้างต้นไม้

ตัวอย่าง 1. ประโยค “The proposals challenge traditional views of accounting practice, based on rights thinking, and adopt concepts from new management theories compatible with the ethics of care”

ผลจากการวิเคราะห์ประโยค คือ

S(NP(The proposals) VP(VP(X(challenge) NP(NP(traditional views) PP(X(of) NP(X(accounting practice) VP(X(,based) PP(X(on) NP(NP(rights) VP(thinking)))))) X(,and) VP(X(adopt) NP(NP(concepts) PP(X(from) NP(NP(X(new management theories) X(compatible) PP(X(with) NP(the ethics))) PP(X(of) NP(care))))))

มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> challenge <NP>

ผลลัพธ์คือ

Subject: The proposals

Object: traditional views of accounting practice, based on rights thinking

Keyword = challenge

จากตัวอย่าง 1 ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ซึ่งกำกับด้วยตัวเลขตามลำดับนั้น แสดงดังรูป 4.9

ตัวอย่าง 2. จากประโยคเดิมเมื่อมีคอนเซ็ปต์แพทเทิร์น คือ

<NP> adopt <NP>

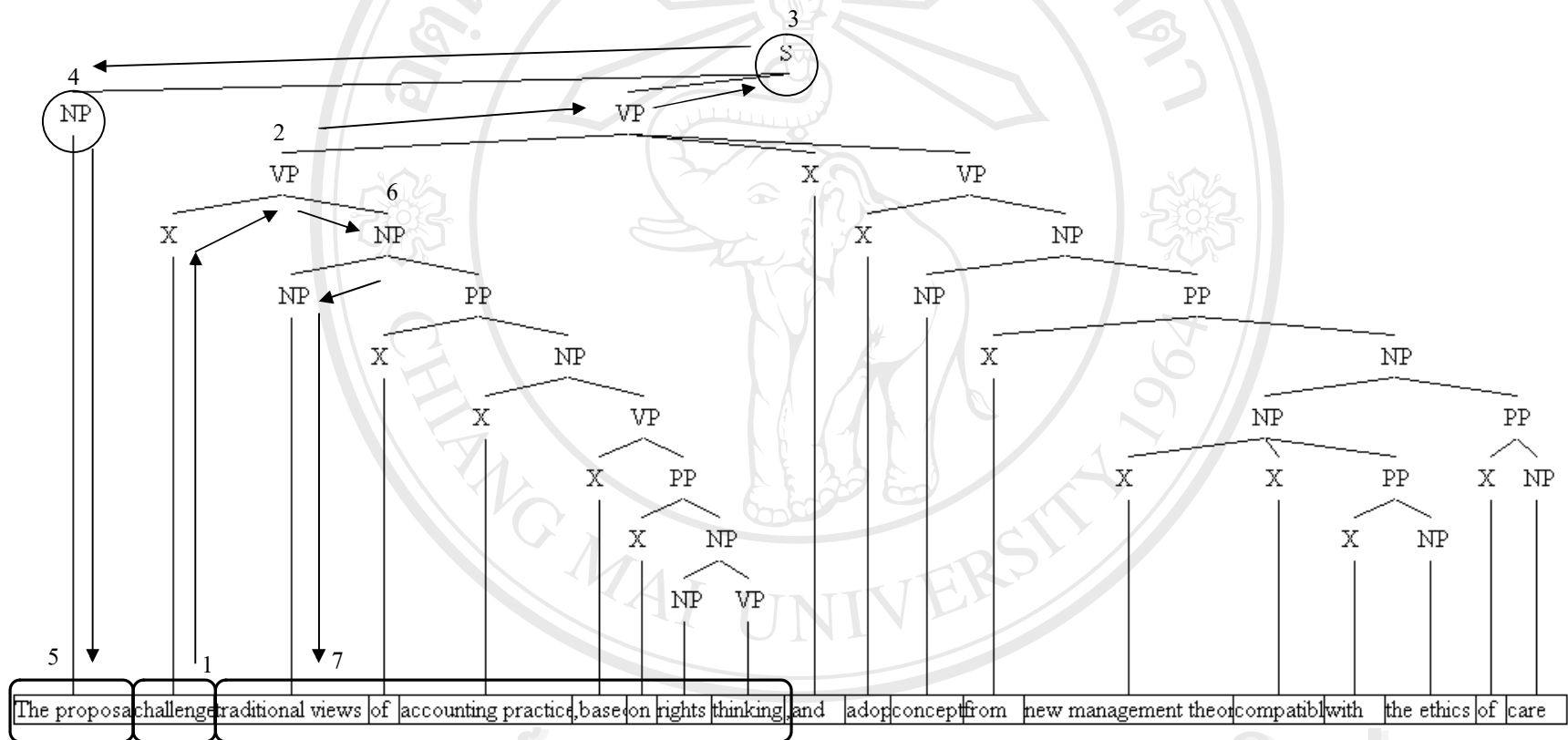
ผลลัพธ์ คือ

Subject: The proposals

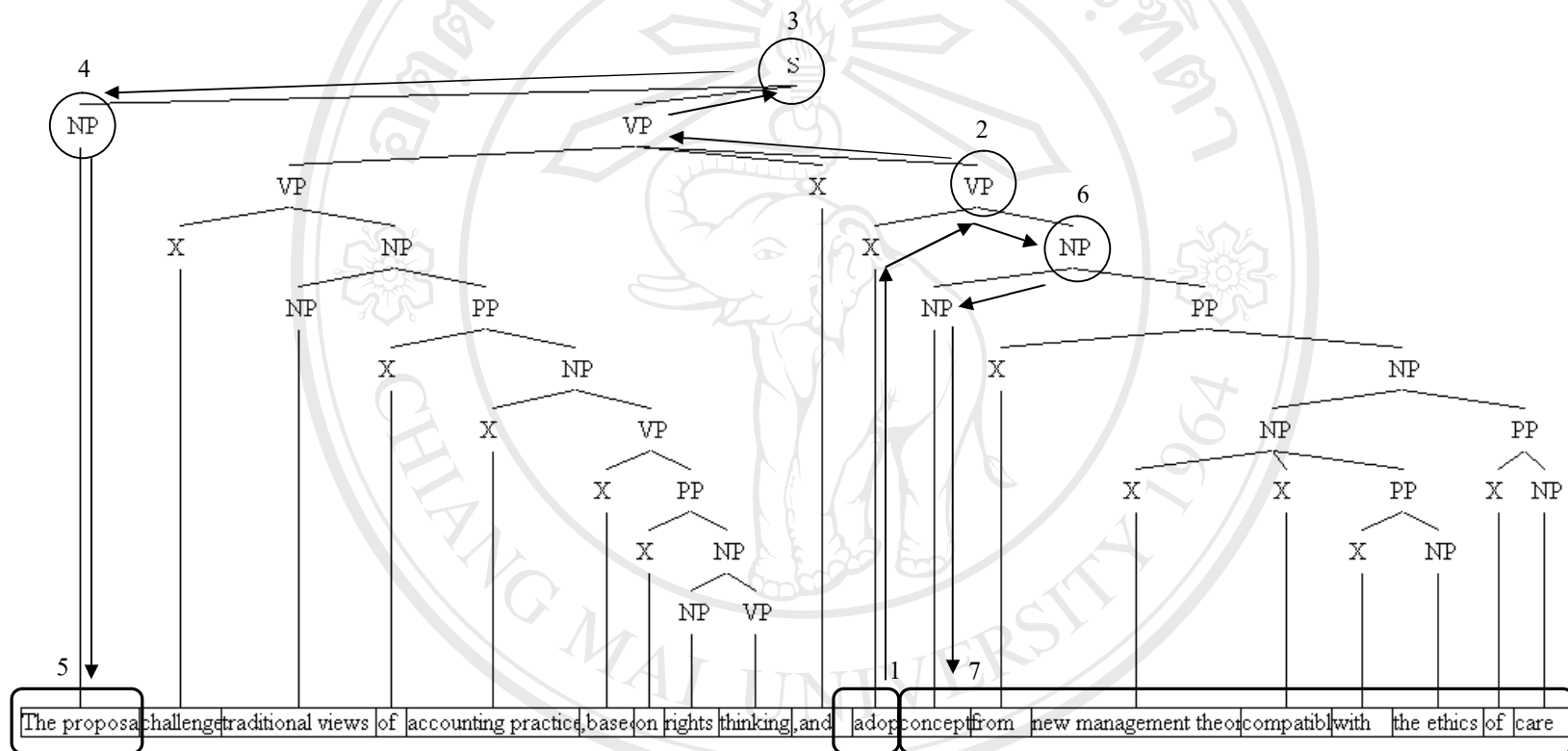
Object: concepts from new management theories compatible with the ethics of care

Keyword = adopt

ประโยคดังกล่าวเป็นประโยคที่ประกอบไปด้วยภาคกริยา 2 ส่วนด้วยกัน คือ “challenge” และ “adopt” ซึ่งมีภาคประธานร่วมกัน พบว่าการคัดแยกข้อความจากโครงสร้างต้นไม้ทำให้สามารถคัดแยกประธานและกรรมที่มีความสัมพันธ์กับคำสำคัญได้อย่างถูกต้อง โดยสามารถท่องเที่ยวส่วนที่ไม่เกี่ยวข้องเพื่อสืบหาประธานของประโยคได้ ซึ่งการคัดแยกโดยวิธีแพทเทิร์นแมชชีนซึ่งไม่สามารถทำได้ ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ลำดับนั้นแสดงดังรูป 4.10



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
 Copyright © by Chiang Mai University
 All rights reserved



รูป 4.10 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 2

ตัวอย่าง 3. ประโยค “Case illustrations underpin the positives and negatives of proactive and reactive business ethics in the marketplace”

ผลจากการวิเคราะห์ประโยค

S (NP (Case illustrations) VP (X (underpin) NP (NP (the positives and negatives) PP (X (of) NP (proactive and reactive business ethics)))) PP (X (in) NP (the marketplace))))

สำหรับประโยคนี้ต้องการจะทำการคัดแยกประธานของประโยคและส่วนที่เป็นบุพบทเพื่อแสดงสถานที่

มีคอนเซ็ปต์แพตเทิร์น คือ

<NP> underpin <PP>

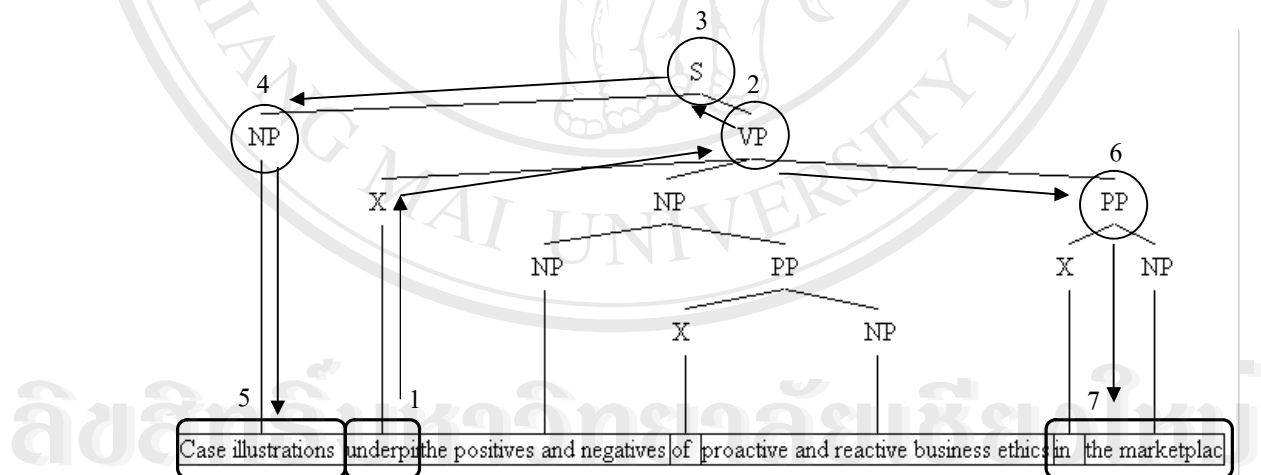
ผลลัพธ์ คือ

Subject: Case illustrations

Preposition: in the marketplace

Keyword = underpin

ผลจากการคัดแยกข้อความและการท่องลงในโครงสร้างต้นไม้ตามลำดับตัวเลข แสดงดังรูป 4.11



รูป 4.11 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 3

จากตัวอย่าง 3 นี้ทำให้เห็นว่าวิธีการที่นำเสนอสามารถท่องไปในโครงสร้างต้นไม้เพื่อทำการคัดแยกส่วนที่ต้องการได้ โดยสามารถท่องข้ามส่วนที่ไม่ต้องการคัดแยกไปได้ตามโครงสร้างต้นไม้ ซึ่งวิธีแพตเทิร์นแมชชีนซึ่งไม่สามารถทำได้

ตัวอย่าง 4. ประโยค “In order to provide some guidance to its members regarding ethical behaviour, the Institute of Internal Auditors has constructed a code of ethics”

ผลจากการวิเคราะห์ประโยค

S (PP (X (In) NP (X (order) S (VP (X (to) VP (X (provide) NP (some guidance) PP (X (to) NP (its members)))))) S (VP (S (X (regarding) NP (ethical behaviour,))) S (NP (X (the Institute) PP (X (of) NP (Internal Auditors))) VP (X (has) VP (X (constructed) NP (NP (a code) PP (X (of) NP (ethics))))))

มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> has constructed <NP>

ผลลัพธ์คือ

Subject: the Institute of Internal Auditors

Object: a code of ethics

Keyword = constructed

ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ตามลำดับตัวเลข แสดงดังรูป 4.12

ตัวอย่าง 5. ประโยค “The results to date of the contemporary organizational ethics movement are not encouraging credible, empirical evidence of more ethical individuals and institutions remains to be discovered”

ผลจากการวิเคราะห์ประโยค

S (NP (NP (NP (The results) PP (to date)) PP (X (of) NP (the contemporary organizational ethics movement))) VP (X (are not) VP (X (encouraging) NP (NP (X (ADJP (credible,)) empirical evidence) PP (X (of) NP (NP (more ethical individuals) X (and) NP (institutions remains)))) S (VP (X (to) VP (X (be) VP (discovered))))))

มีคอนเซ็ปต์แพทเทิร์น คือ

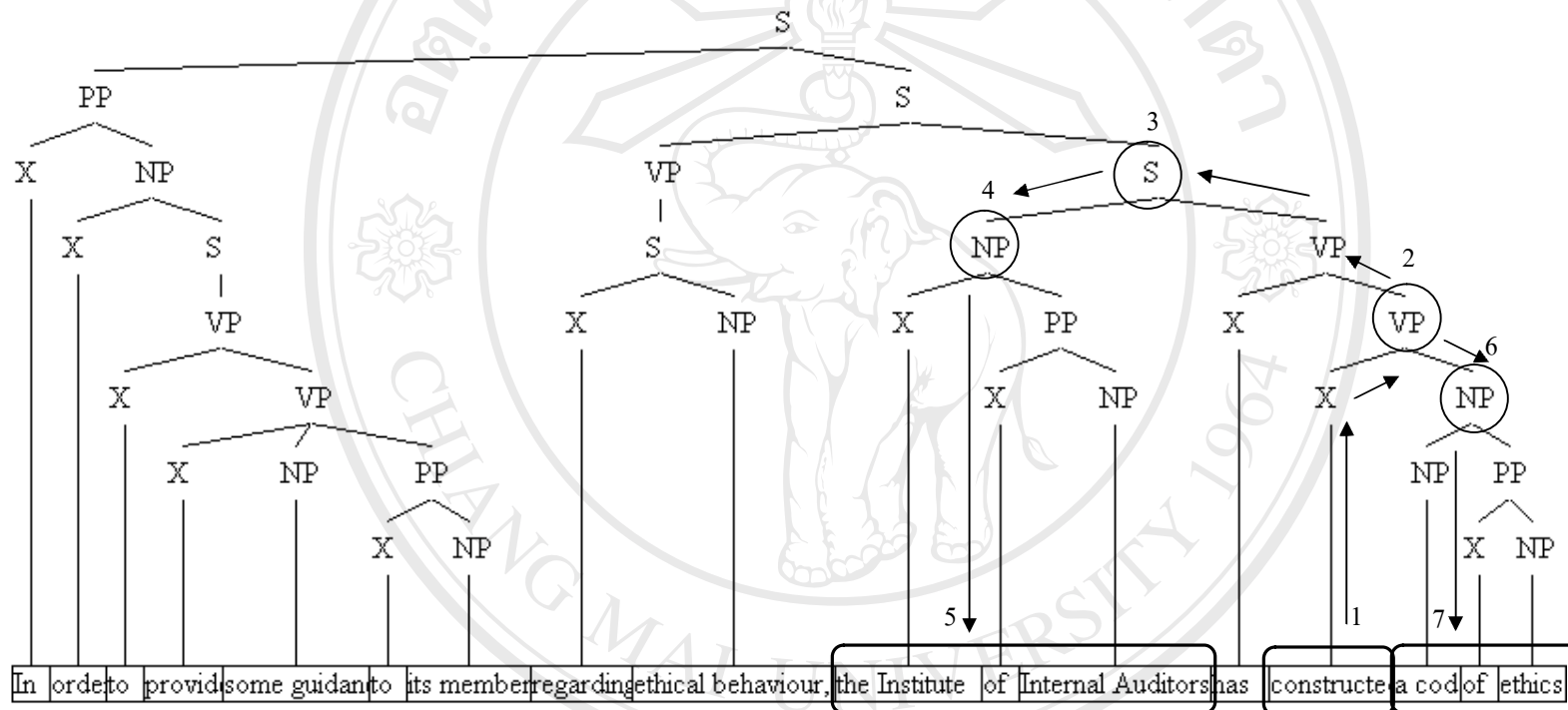
<NP> are not encouraging <NP>

ผลลัพธ์ คือ

Subject: The results to date of the contemporary organizational ethics movement

Object: empirical evidence of more ethical individuals and institutions remains

Keyword = encouraging



รูป 4.12 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 4

ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ตามลำดับตัวเลข แสดงดังรูป 4.13 สามารถคัดแยกกรรมโดยละเว้น “credible” ซึ่งเป็นส่วนขยายของนามวลีที่ทำหน้าที่เป็นกรรมได้

ตัวอย่าง 6. ประโยค “An analysis of real-life ethical situations leads to the formulation of three models of corporate codes of conduct, and to conclusions as to the pivotal role the ethical dimension can play in corporate life”

ผลจากการวิเคราะห์ประโยค

S (NP (NP (An analysis) PP (X (of) NP (real-life ethical situations))) VP (X (leads) PP (PP (X (to) NP (NP (NP (the formulation) PP (X (of) NP (three models))) PP (X (of) NP (NP (corporate codes) PP (X (of) NP (conduct)))))) X (, and) PP (X (to) NP (conclusions))) PP (X (as to) NP (NP (the pivotal role) SBAR (S (NP (the ethical dimension) VP (X (can) VP (X (play) PP (X (in) NP (corporate life))))))))))

มีคอนเซ็ปต์แพตเทิร์น คือ

<NP> leads <PP>

ผลลัพธ์ คือ

Subject: An analysis of real-life ethical situations

Preposition: to the formulation of three models of corporate codes of conduct, and to conclusions

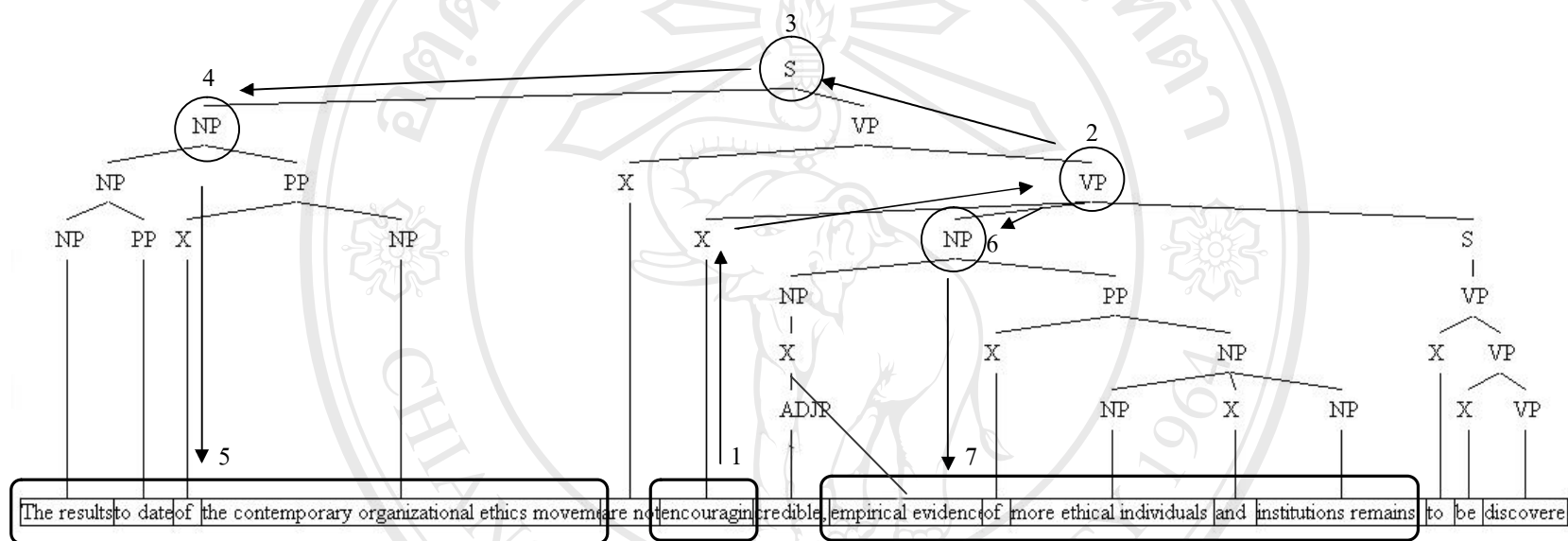
Keyword = leads

ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ตามลำดับตัวเลข แสดงดังรูป 4.14

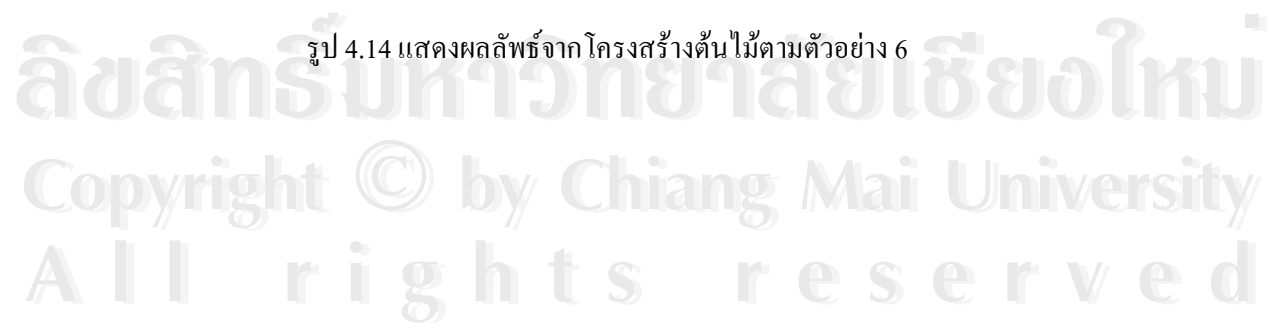
ตัวอย่าง 7. ประโยค “The ethics of rights or the separative model has dominated Western thought since the Enlightenment and the ethics of care was developed as a feminist critique seeking to rebalance our basic thought structure”

ผลจากการวิเคราะห์ประโยค

S (S (NP (NP (The ethics) PP (X (of) NP (rights))) X (or) NP (the separative model) VP (X (has) NP (dominated Western thought) PP (X (since) NP (the Enlightenment)))) X (and) S (NP (NP (the ethics) PP (X (of) NP (care))) VP (X (was) VP (X (developed) PP (X (as) NP (NP (a feminist critique) VP (X (seeking) S (VP (X (to) VP (X (rebalance) NP (our basic thought structure))))))))))



รูป 4.13 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 5



มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> has dominated <NP>

จากตัวอย่าง 7 สามารถแสดงโครงสร้างต้นไม้ดังรูป 4.15 ซึ่งจากโครงสร้างต้นไม้ดังกล่าว พบว่า การทำการสืบค้นเพื่อค้นหาทริกเกอร์เวิร์ดนั้นทำให้ไม่สามารถทำการคัดแยกข้อความที่ต้องการได้ เนื่องจากพาสเซอร์ทำการวิเคราะห์ว่าทริกเกอร์เวิร์ด “dominated” นั้นเป็นส่วนหนึ่งของนามวลี การสืบค้นจึงไม่พบทริกเกอร์เวิร์ด

ตัวอย่าง 8. ประโยค “The ultimate aim of this ethics training program is to advance global democratic deliberation and decision making in both private and public organizations”
ผลจากการวิเคราะห์ประโยค

S(NP(NP(The ultimate aim) PP(X(of) NP(this ethics training program)))
VP(X(is)S(VP(X(to) VP(X(advance) NP(NP(global democratic deliberation) X(and)
NP(decision making)) PP(X(in) NP(X(both private and public organizations)))))))))

มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> is to advance <NP>

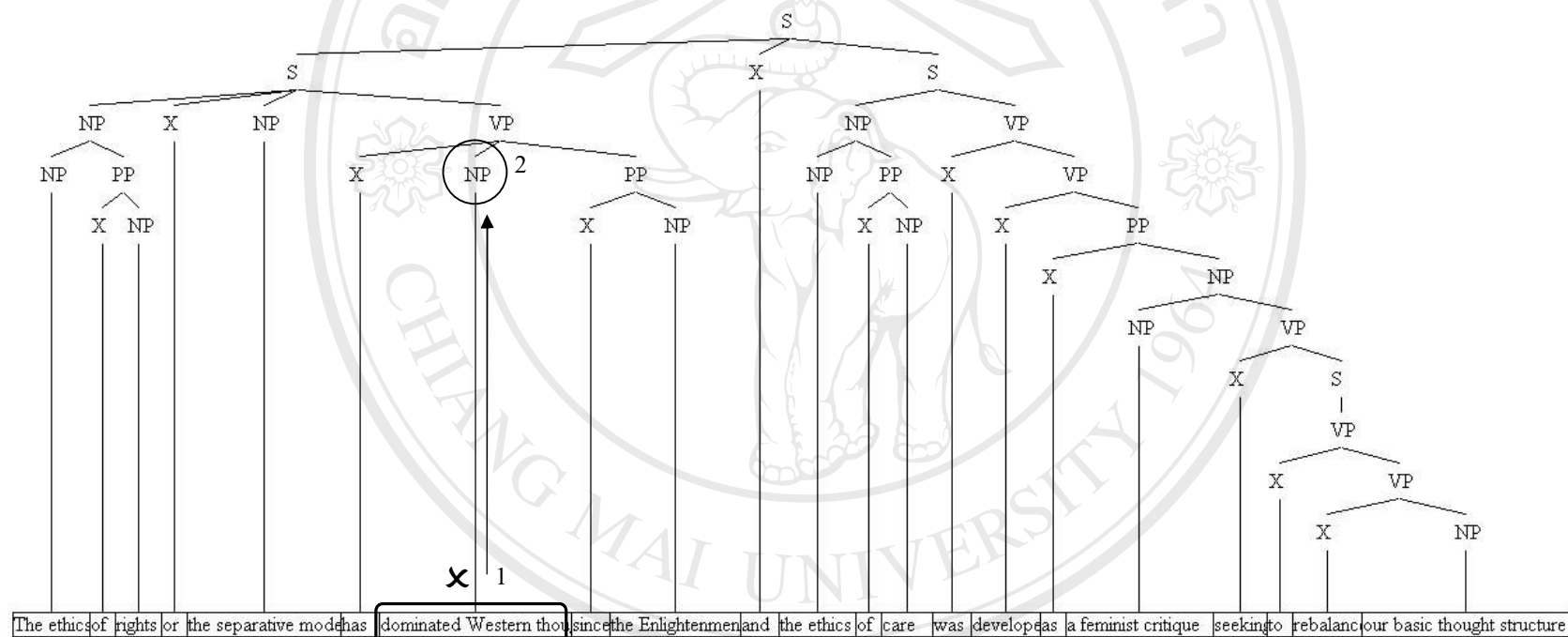
ผลลัพธ์ คือ

Subject: global democratic deliberation and decision making

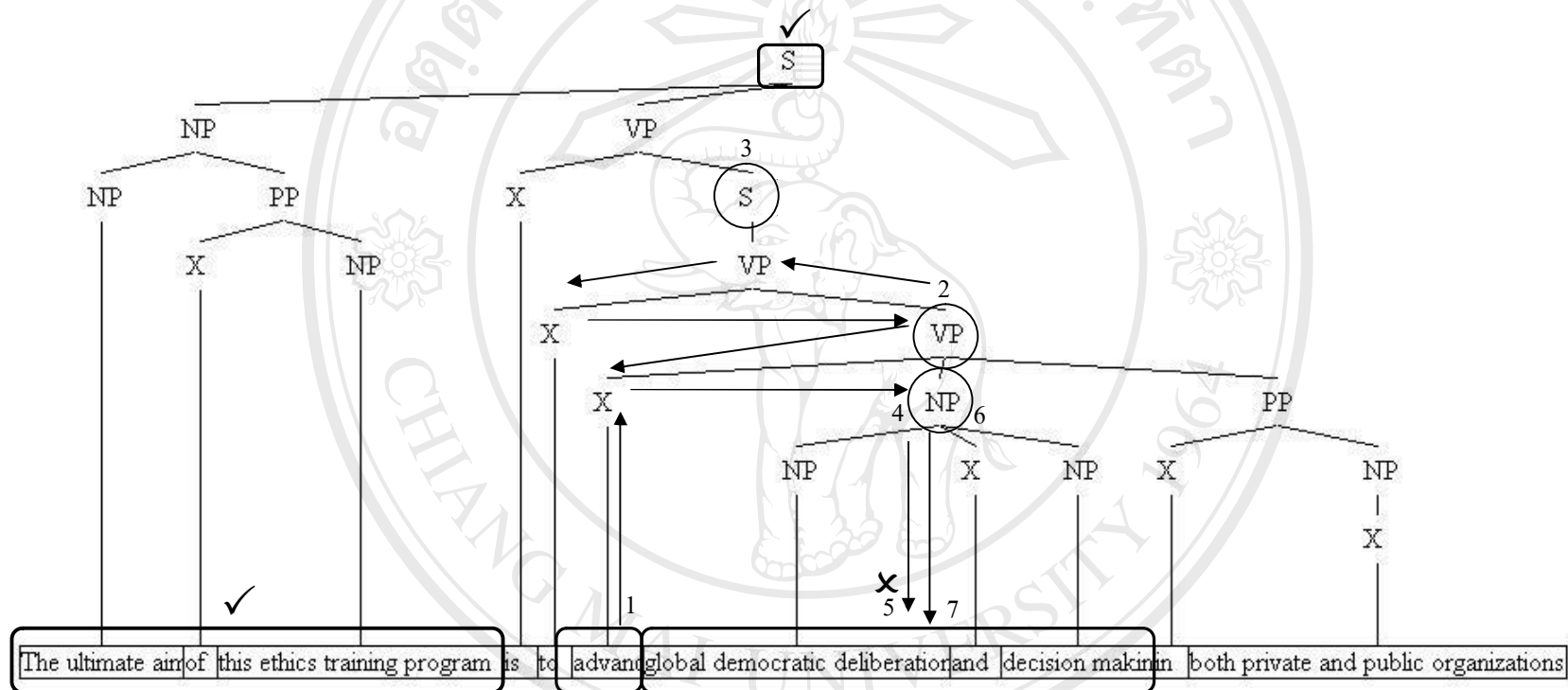
Object: global democratic deliberation and decision making

Keyword = advance

ซึ่งจากประโยคในตัวอย่างที่ 8 นี้ไม่สามารถทำการคัดแยกประธานมาได้อย่างถูกต้อง เนื่องจากความผิดพลาดในการท่องลงไปโครงสร้างต้นไม้เพื่อค้นหาประธาน เมื่อเจอโหนด “S” ให้ทำการสืบค้นลงมาเพื่อหามวลี “NP” แต่ผลจากการวิเคราะห์โครงสร้างนั้นพบว่าพาสเซอร์วิเคราะห์ว่ามีอนุประโยคซ้อนอยู่ในประโยคหลัก ซึ่งอนุประโยคนั้นกำกับด้วยโหนด “S” และเริ่มต้นท่องลงไปหามวลีเมื่อเจอโหนด “NP” จึงคัดแยกมาซึ่งได้ผลผิดพลาดไป โดยที่ถูกต้องนั้นจะต้องท่องขึ้นไปจนกระทั่งพบโหนด “S” ที่เป็นรูทโหนดจึงจะสามารถคัดแยกประธานได้อย่างถูกต้อง ผลจากการคัดแยกจากข้อความและการท่องลงไปโครงสร้างต้นไม้ตามลำดับตัวเลข แสดงดังรูป



รูป 4.15 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 7



รูป 4.16 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 8

ตัวอย่าง 9 และ 10 นี้ เป็นตัวอย่างที่แสดงให้เห็นว่ารูปแบบของประโยคนั้นมีผลต่อประสิทธิภาพของการวิเคราะห์โครงสร้างของประโยค สำหรับประโยคในสาขาไวยากรณ์นั้น มักจะเป็นประโยคที่มีความซับซ้อน แม้ว่าจะทำการอ่านและตีความหมายโดยมนุษย์ก็ยังคงเกิดความสับสนได้ เนื่องจากลักษณะของประโยคไม่ใช่ประโยคความเดียว แต่เป็นประโยคที่มีการอธิบายความ มีคำกริยาซึ่งแสดงปฏิกิริยาของสารต่าง ๆ หลายคำด้วยกันในหนึ่งประโยค และมีลักษณะเป็นประโยคซ้อนประโยค มีการใช้คำเชื่อมมากกว่าหนึ่งคำในประโยค มีการใช้ศัพท์เฉพาะซึ่งอาจทำให้พาสเซอร์เดาหน้าที่ของคำผิดพลาด ดังนั้นเมื่อทำการวิเคราะห์โครงสร้างของประโยคโดยลิงก์แกรมมาพาสเซอร์แล้ว จึงทำให้เกิดความผิดพลาดและมีผลให้การคัดแยกผิดพลาดด้วยเช่นกัน ดังตัวอย่างต่อไปนี้

ตัวอย่าง 9. ประโยค “Removing the intron from TUB1 in cef1-13 cells boosts TUB1

mRNA and alpha-tubulin expression to near wild-type levels and restores microtubule stability in the cef1-13 mutant”

ผลจากการวิเคราะห์ประโยค

S(NP(X(Removing) NP(the intron) PP(X(from) NP(TUB1)) PP(X(in) NP(cef1-13 cells))) VP(VP(X(boosts) NP(TUB1 mRNA) X(and) NP(alpha-tubulin expression)S(VP(X(to) VP(X(near) NP(wild-type levels)))) X(and) VP(X(restores) NP(NP(microtubule stability) PP(X(in) NP(the cef1-13 mutant))))))

มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> boosts <NP>

ผลลัพธ์ คือ

Subject: Removing the intron from TUB1 in cef1-13 cells

Object: TUB1 mRNA

Keyword = boosts

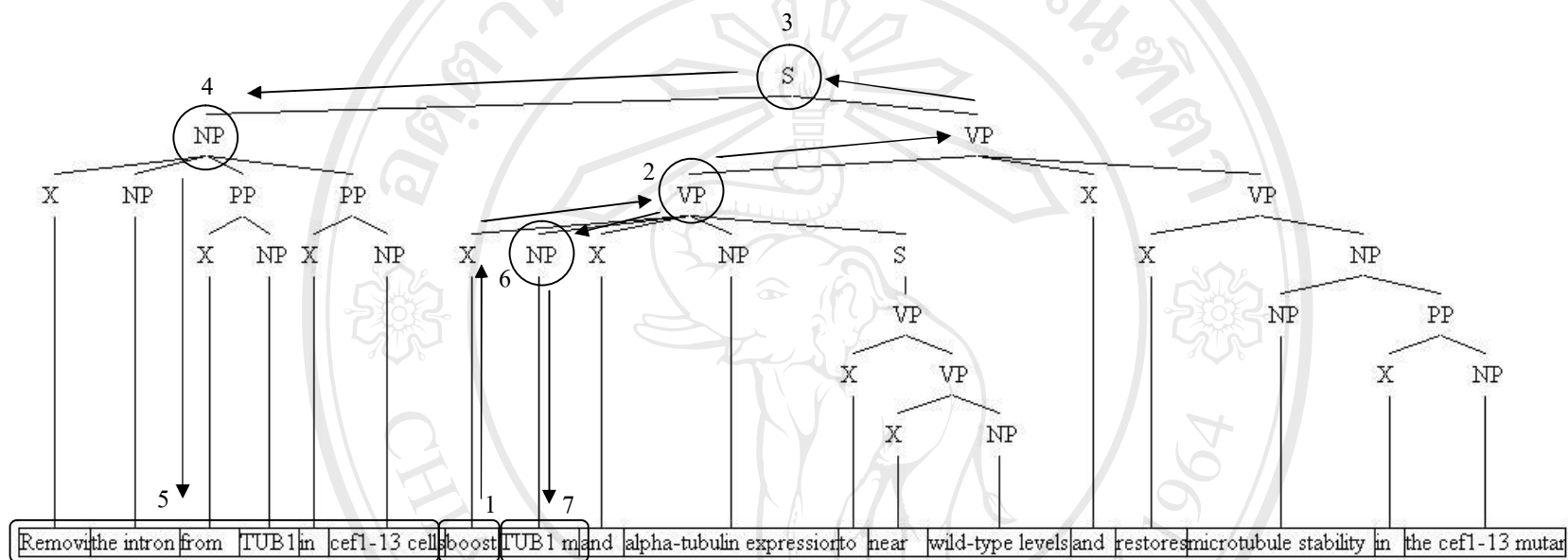
ผลที่ได้จากการคัดแยกในส่วนของนามวลีที่ทำหน้าที่เป็นกรรมของประโยคในตัวอย่างนี้

พบว่าไม่สามารถคัดแยกได้อย่างสมบูรณ์ เนื่องจากมีการใช้คำเชื่อม “and” ในส่วนที่ต้องการทำการคัดแยกข้อความ และพาสเซอร์ทำการวิเคราะห์ในส่วนของคำนามวลีที่เชื่อมกันด้วย “and” แยก

ออกเป็นนามวลีสองส่วนด้วยกัน ทำให้ไม่สามารถคัดแยกข้อความที่ต้องการมาได้อย่างครบถ้วน

ผลจากการคัดแยกข้อความและการท่องเที่ยวไปในโครงสร้างต้นไม้ตามลำดับตัวเลข สามารถแสดงได้

ดังรูป 4.17



รูป 4.17 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 9

ตัวอย่าง 10. ประโยค “As a result, cef1-13 tub1Deltai cells progress through mitosis and their cell cycle arrest phenotype is alleviated”

ผลจากการวิเคราะห์ประโยค

S(PP(X(As) NP(NP(a result) X(,) NP(cef1-13 tub1Deltai)))S(NP(cells)
VP(X(progress) PP(X(through) NP(mitosis))) X(and)S(NP(their cell cycle arrest phenotype)
VP(X(is) VP(alleviated)))))

มีคอนเซ็ปต์แพทเทิร์น คือ

<NP> progress <PP>

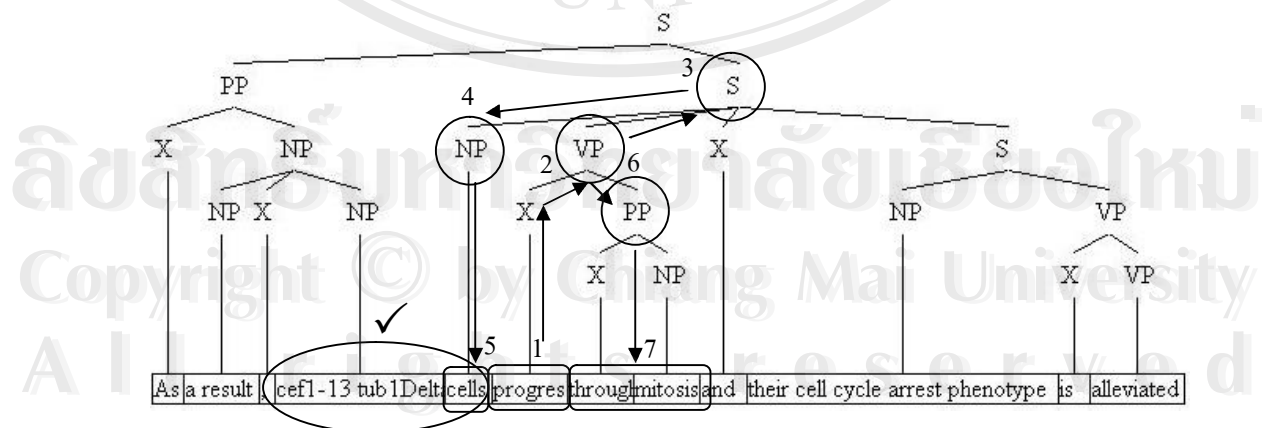
ผลลัพธ์ คือ

Subject: cells

Preposition: through mitosis

Keyword = progress

ผลที่ได้จากการคัดแยกในส่วนของนามวลีที่ทำหน้าที่เป็นประธานของประโยคในตัวอย่างนี้ ไม่สามารถคัดแยกได้อย่างสมบูรณ์ เนื่องจากในประโยคมีการใช้คำศัพท์เฉพาะทำให้พาสเซอร์ระบุหน้าที่ของคำผิดพลาดและมีผลต่อชนิดของลิงก์ที่เกิดขึ้น ดังนั้นเมื่อชนิดของลิงก์ที่ได้ไม่ถูกต้องจึงมีผลทำให้การวิเคราะห์โครงสร้างของประโยคผิดพลาดไปด้วย ดังนั้นเมื่อการวิเคราะห์ห้วงค์ประกอบของวลีในประโยคผิดพลาดไป จึงไม่สามารถคัดแยกข้อความที่ต้องการมาได้อย่างครบถ้วน ผลจากการคัดแยกข้อความข้อความและการท่องลงไปในโครงสร้างต้นไม้ตามลำดับตัวเลข สามารถแสดงได้ดังรูป 4.18



รูป 4.18 แสดงผลลัพธ์จากโครงสร้างต้นไม้ตามตัวอย่าง 10

สำหรับผลลัพธ์ที่ได้จากการทดลองกับบทความในสาขาสังคมวิทยา และบทความใน
สาขาไบโอเคมีนั้น จะกล่าวถึงในบทถัดไป โดยการทดลองนี้จะทำการประเมินผลด้วยค่าพรีซิชั่น
และค่ารีคอลซึ่งเป็นค่ามาตรฐานที่ใช้ในการประเมินระบบการคัดแยกข้อมูล



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright © by Chiang Mai University
All rights reserved