TE9303

# การแปลภาษาด้วยคอมพิวเตอร์
# Machine Translation

โครงการพัฒนาระบบเครื่องแปลภาษาสำหรับภาษาในเอเชีย




418.02
ก154
2536
ฉ.2

าคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
the International Cooperation for Computerization
CICC

# การแปลภาษาด้วยคอมพิวเตอร์
# Machine Translation

## โครงการพัฒนาระบบเครื่องแปลภาษาสำหรับภาษาในเอเชีย


ผู้ร่วมโครงการ:

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์
สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี

ภาควิชาอิเล็กทรอนิกส์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง



b 014022

NSTDA สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

000034
2/9/01
*B17268

การแปลภาษาด้วยคอมพิวเตอร์

Machine Translation



QR Code by
RFID Lab NECTEC, STKS

จัดพิมพ์โดย:

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC)

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (NSTDA)

กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

พิมพ์ครั้งที่ 1

มีนาคม 2536

NSTDA-E-93-07B01

สวท.
ศ0-8
0109
2536
ฉ.2

418.02
ก754
2536
ฉ.2

สวท 000755 ศ0

# คำนำ

หนังสือเล่มนี้เป็นการรวบรวมบทความด้านการแปลภาษาด้วยเครื่องคอมพิวเตอร์ที่เป็นโครงการความร่วมมือระหว่างไทย-ญี่ปุ่น โดยมีจุดมุ่งหมายเพื่อนำเสนอผลงานต่อสาธารณะชน

บทความในหนังสือเป็นเนื้อหาทางวิชาการที่เกี่ยวข้องกับการแปลภาษาโดยตรง ผลงานดังกล่าวนี้แม้ว่าจะเป็นเพียงบางส่วนไม่ใช่ผลงานทั้งหมด แต่ก็ถือได้ว่าเป็นส่วนสำคัญของโครงการฯ

เนื่องจากระยะเวลาในการรวบรวมเรื่องราวต่างๆของโครงการฯมีไม่มากนัก ดังนั้นการนำเสนอเนื้อหาจึงเป็นเพียงหลักการกว้างๆเท่านั้น สำหรับรายละเอียดสามารถหาดูได้จากรายงานความก้าวหน้าของแต่ละโครงการย่อย

หวังว่าหนังสือเล่มนี้จะให้ประโยชน์ทางด้านวิชาการกับผู้ที่สนใจ

บุญเจริญ ศิริเนาวกุล

บรรณาธิการ

มีนาคม 2536

# สารบัญ


โครงการแปลภาษาด้วยคอมพิวเตอร์ 7

การแสดงรูปแทนภาษากลางสำหรับภาษาไทย 11

ระบบการวิเคราะห์ภาษาไทยเพื่อการแปลภาษาด้วยคอมพิวเตอร์ 31

การตัดคำไทยในระบบแปลภาษา 50

การสังเคราะห์ภาษาไทยจากภาษากลางในระบบการแปลหลายหลากภาษา 57

โครงการพัฒนาพจนานุกรมเพื่อการแปลภาษาด้วยคอมพิวเตอร์ 75

การสร้างระบบรวมเพื่อการแปลภาษาด้วยคอมพิวเตอร์ 103

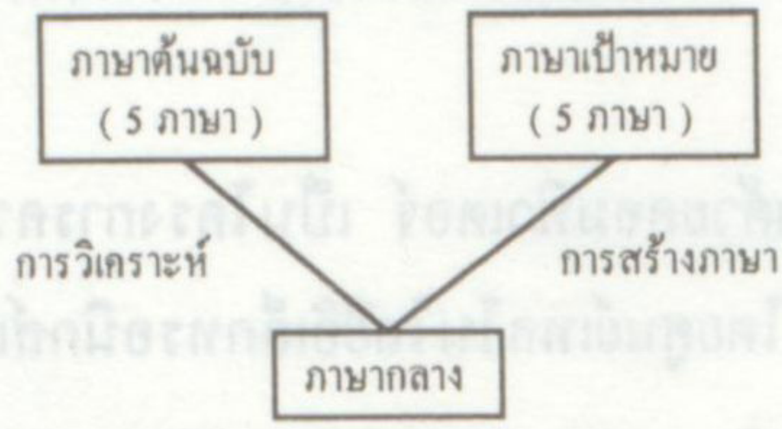
อักขรวิธีภาษาไทยสำหรับระบบX Window 116

## โครงการแปลภาษาด้วยคอมพิวเตอร์

โครงการแปลภาษาด้วยคอมพิวเตอร์ เป็นโครงการความร่วมมือระหว่างรัฐบาลไทย และรัฐบาลญี่ปุ่น โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (National Electronics and Computer Technology Center : NECTEC) กระทรวงวิทยาศาสตร์เทคโนโลยีและสิ่งแวดล้อม เป็นหน่วยงานที่ดูแลโครงการฝ่ายไทย และ Center of International Cooperation for Computerization (CICC) ซึ่งเป็นหน่วยงานหนึ่งของ Ministry of International Trade and Industry (MITI) เป็นหน่วยงานที่ดูแลโครงการฝ่ายญี่ปุ่น โครงการนี้ มีกำหนดเวลาการดำเนินงานวิจัย 6 ปี ตั้งแต่ปี พ.ศ. 2530 ถึง พ.ศ. 2535 แต่เนื่องจากการดำเนินการวิจัยในระยะแรก มีอุปสรรคเกี่ยวกับงบประมาณและจำนวนนักวิจัย ไม่สามารถดำเนินการวิจัยโดยเต็มที่ได้ กำหนดเวลาการสิ้นสุดโครงการจึงต้องเลื่อนไปเป็นปี พ.ศ.2537

วัตถุประสงค์ของโครงการนี้ คือการวิจัยการแปลภาษาด้วยคอมพิวเตอร์แบบหลากภาษา (Multi-Lingual Machine Translation) โดยนักวิจัยฝ่ายไทยสร้างฐานความรู้ในการวิเคราะห์ภาษาไทยเพื่อให้คอมพิวเตอร์ประมวลผลเป็นภาษากลาง(interlingua) และนำภาษากลางมาสร้างเป็นภาษาไทยอีกทอดหนึ่ง สำหรับนักวิจัยฝ่ายญี่ปุ่นนั้น จะทำการวิจัยเช่นเดียวกันนี้ในภาษาญี่ปุ่น เมื่อ นำระบบทั้งสองรวมเข้าด้วยกันแล้ว คอมพิวเตอร์จะสามารถแปลจากภาษาไทยหรือญี่ปุ่น เป็นภาษาไทยหรือญี่ปุ่นได้

นอกจากความร่วมมือกับประเทศไทยแล้ว CICC ยังได้ร่วมวิจัยกับประเทศจีน อินโดนีเซีย และมาเลเซีย ในลักษณะคู่ขนานไปด้วย ซึ่งเมื่อเสร็จสิ้นโครงการแล้วจะสามารถใช้คอมพิวเตอร์แปลภาษาไทย จีน ญี่ปุ่น อินโดนีเซีย หรือมาเลเซียออกมาเป็นภาษาใดภาษาหนึ่งใน 5 ภาษานี้ได้ การทำงานของโครงการแปลภาษาด้วยคอมพิวเตอร์ในระบบหลากภาษานี้อาจสรุปเป็นแผนภูมิง่ายๆ ดังนี้




รูปที่ 1 การแปลภาษาด้วยเครื่องคอมพิวเตอร์ระบบหลากภาษา

ในด้านงานวิจัยของฝ่ายไทยนั้น สามารถอธิบายโดยย่อดังนี้

1. ป้อนข้อมูลด้วย Thai Word Processor โดยผลที่ได้จากข้อมูลนี้จะเป็นภาษาต้นฉบับ (Source Language) ให้ตัววิเคราะห์ภาษา (Analyser)

2. ตัววิเคราะห์ภาษา จะมีฐานความรู้ทางด้านกฎเกณฑ์ภาษาไทยตามทฤษฎีภาษาศาสตร์เพื่อวิเคราะห์ภาษาไทยโดยอาศัยข้อมูลจากพจนานุกรม และทำการสร้างภาษาต้นฉบับเป็นโครงสร้างภาษากลาง (Interlingua Structure)

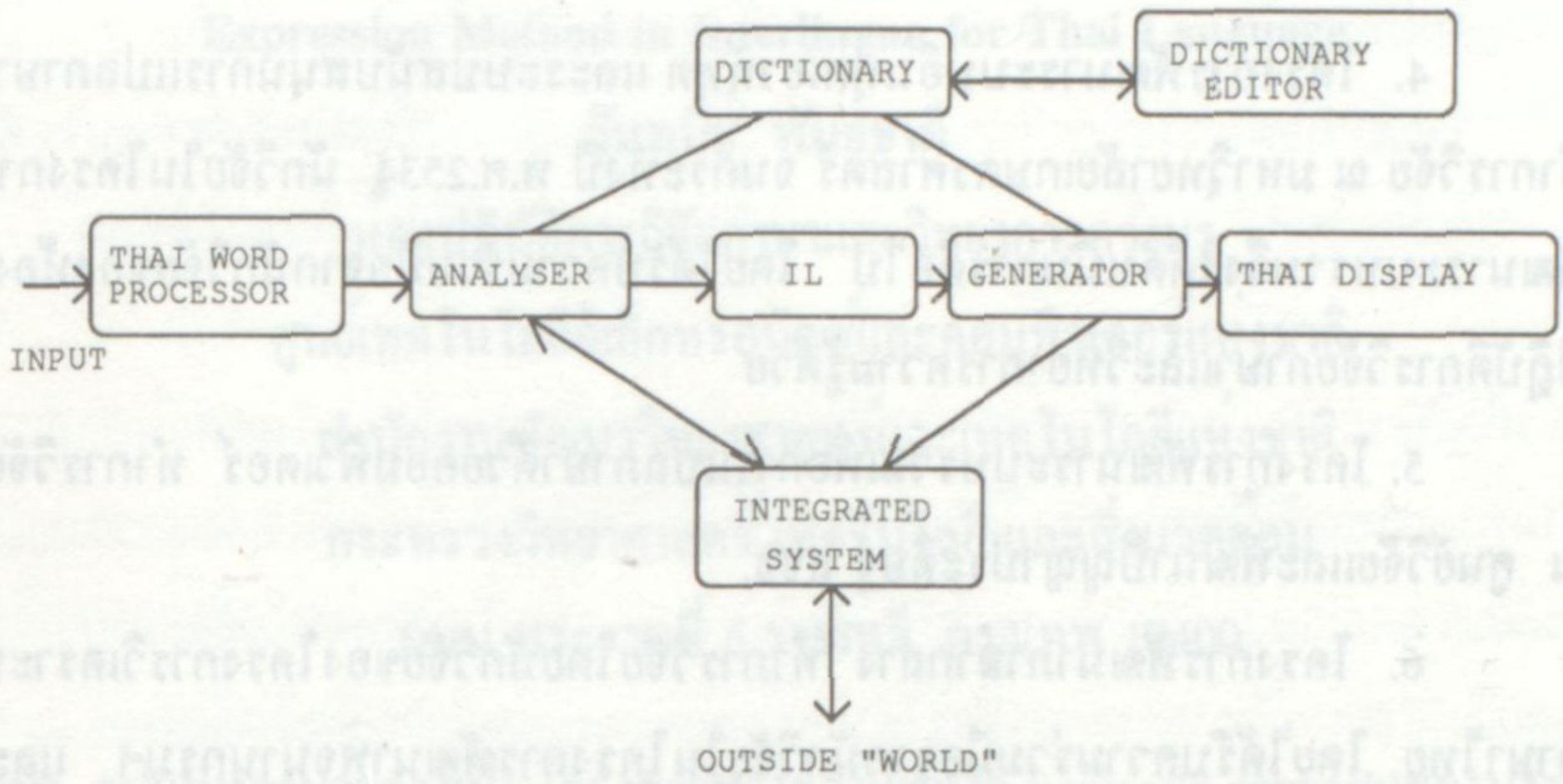
3. ตัวสร้างภาษาเป้าหมาย (Generator) ทำการถอดความจากโครงสร้างภาษากลางให้เป็นภาษาไทย โดยอาศัยกฎเกณฑ์การสร้างภาษาไทยจากทฤษฎีภาษาศาสตร์และข้อมูลจากพจนานุกรม

4. ผลของการสร้างภาษาเป้าหมายนี้ จะแสดงผลเป็นภาษาไทยบนเครื่องคอมพิวเตอร์ด้วยการใช้โปรแกรมแสดงผลภาษาไทยบนหน้าจอ

5. การทำงานทั้งหมดนี้ มีการจัดลำดับขั้นตอนของแต่ละโมดูลต่างกันซึ่งจะมีระบบรวม (Integrated System) เป็นผู้ควบคุม

6. พจนานุกรมจะมีการแก้ไข เพิ่มเติมข้อมูลได้ด้วยการใช้ Dictionary Editor

งานดังกล่าวอาจสรุปเป็นรูปแผนภูมิดังนี้




รูปที่ 2 การทำงานของโครงการแปลภาษาด้วยคอมพิวเตอร์

เนื่องจากการทำงานเป็นโมดูลต่างๆ ดังกล่าวนี้ โครงการแปลภาษาด้วยคอมพิวเตอร์ จึงแบ่งเป็นโครงการย่อย 6 โครงการ ผู้ทำการวิจัยเป็นอาจารย์ และนักวิจัย ทั้งทางด้านอักษรศาสตร์ และคอมพิวเตอร์จากหน่วยงานวิจัย ของมหาวิทยาลัย และของรัฐบาล ซึ่งอาจกล่าวโดยละเอียดดังนี้

1. โครงการพัฒนาพจนานุกรมเพื่อการแปลภาษาด้วยคอมพิวเตอร์ ทำการวิจัย ณ ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์ สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี (สจธ.)

2. การวิเคราะห์ภาษาไทยเพื่อการแปลภาษาด้วยคอมพิวเตอร์ ทำการวิจัย ณ คณะอักษรศาสตร์จุฬาลงกรณ์มหาวิทยาลัย จนถึงต้นปี พ.ศ.2533 จากนั้นสถาบันเทคโนโลยีสังคม (เกริก) ร่วมกับศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์ สจธ.ดำเนินการต่อไป จนกระทั่งต้นปี 2535 ซึ่งศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ได้ตั้งห้องปฏิบัติการวิจัยขึ้นใหม่คือห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ (Linguistics and Knowledge Science Laboratory: LINKS) ห้องปฏิบัติการนี้จึงได้รับงานส่วนนี้ไปดำเนินการต่อไป

3. โครงการสร้างภาษาเป้าหมายเพื่อการแปลภาษาด้วยคอมพิวเตอร์ ทำการวิจัย ณ ห้องปฏิบัติการแปลภาษาด้วยคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

4. โครงการพัฒนาระบบอินพุตเอาท์พุต และระบบสนับสนุนการแปลภาษา ทำการวิจัย ณ มหาวิทยาลัยเกษตรศาสตร์ จนกระทั่งปี พ.ศ.2534 นักวิจัยในโครงการพัฒนาระบบรวมจึงได้ดำเนินการต่อไป โดยได้รับความร่วมมือจากนักวิจัยของห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ด้วย

5. โครงการพัฒนาระบบรวมเพื่อการแปลภาษาด้วยคอมพิวเตอร์ ทำการวิจัย ณ ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์ สจธ.

6. โครงการพัฒนาภาษากลาง ทำการวิจัยโดยนักวิจัยของโครงการวิเคราะห์ภาษาไทย โดยได้รับความร่วมมือจากนักวิจัยในโครงการพัฒนาพจนานุกรมฯ และโครงการสร้างภาษาเป้าหมายฯ ปัจจุบันนี้ดำเนินงานวิจัย ณ ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้

สำหรับงบประมาณการทำการวิจัยในโครงการนี้ทั้งหมดได้รับความสนันสนุนทั้งจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ และ CICCโดยแต่ละปีจะได้รับเงินอุดหนุนการวิจัยรวมทั้งสิ้นประมาณ 8 ล้านบาท สำหรับทั้ง 6 โครงการ โดยแต่ละโครงการมีหัวหน้าโครงการรับผิดชอบเกี่ยวกับการใช้เงินเพื่อดำเนินงานวิจัยให้บรรลุวัตถุประสงค์ของงานในแต่ละปี

# การแสดงรูปแทนภาษากลางสำหรับภาษาไทย

# Expression Method in Interlingua for Thai Language

**วันทนีย์ พันธชาติ**

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

กระทรวงวิทยาศาสตร์ เทคโนโลยี และสิ่งแวดล้อม

ถนน พระรามที่ 6 ราชเทวี กรุงเทพ 10400

**1. บทนำ** รูปแทนภาษากลางเป็นโครงข่ายทางความหมาย (semantic network)ของประโยค รูปแบบและข้อกำหนดภาษากลางออกแบบโดย อูชิดะ(Uchida : 1991) และพัฒนาโดยนักวิจัยจาก 5 ประเทศที่เข้าร่วมโครงการเครื่องแปลภาษาสำหรับภาษาในเอเชีย ซึ่งได้แก่ ประเทศญี่ปุ่น จีน มาเลเซีย อินโดนีเซีย และไทย ทุกประเทศที่เข้าร่วมโครงการใช้ข้อกำหนดนี้เพื่อแสดงรูปแทนทางความหมาย(semantic representation) สำหรับเครื่องแปลภาษาของประเทศนั้น ๆ ในบทความนี้ผู้วิจัยขอเสนอการแทนรูปประโยคภาษาไทยด้วยโครงข่ายทางความหมายตามโครงสร้างของภาษากลาง

**2. โครงสร้างของรูปแทนภาษากลาง** รูปแทนภาษากลางแสดงความสัมพันธ์ของมโนทัศน์ 2 มโนทัศน์ด้วยการกสัมพันธ์ และคุณลักษณ์ในรูปของ tuple form ดังนี้

<มโนทัศน์> ::= <มโนทัศน์หลัก> | <มโนทัศน์ประกอบ>

<มโนทัศน์ประกอบ>: := {<ความสัมพันธ์แบบไบนารี>[ความสัมพันธ์แบบไบนารี> ...]
[ <ความสัมพันธ์แบบยูนารี>} |
{ <ความสัมพันธ์แบบยูนารี>[ความสัมพันธ์แบบไบนารี>...]
[ <ความสัมพันธ์แบบยูนารี>]}
( <คุณลักษณ์>,รูปว่าง,<มโนทัศน์>,<ความแน่นอน> )

<ความสัมพันธ์แบบไบนารี> ::= ( <ความสัมพันธ์>,<มโนทัศน์>,<มโนทัศน์> )

<ความสัมพันธ์แบบยูนารี>  ::= ( <คุณลักษณ์>,<มโนทัศน์>,รูปว่าง)

<มโนทัศน์หลัก>  ::= รูปที่แทนมโนทัศน์

<ความสัมพันธ์>  ::= รูปสำหรับแทนความสัมพันธ์ระหว่างมโนทัศน์

<คุณลักษณ์>  ::= รูปสำหรับแทนคุณลักษณ์ของมโนทัศน์

**2.1 มโนทัศน์หลัก** มโนทัศน์หลักจะอธิบายมโนทัศน์ในประโยค มโนทัศน์หลักที่ใช้ในภาษากลางนี้เป็นมโนทัศน์หลักที่พัฒนาขึ้นโดย EDR (Japan Electronic Dictionary Research Institute,Ltd.) ประเทศญี่ปุ่น ที่จริงการกำหนดมโนทัศน์ของแต่ละภาษาเป็นผลมาจากการที่เจ้าของภาษารับรู้โครงสร้างของการแยกแยะ (conceptualization) สิ่งต่าง ๆในโลก ซึ่งแต่ละชาติและแต่ละภาษาจะต่างกันออกไป เพราะวัฒนธรรมและการรับรู้สิ่งต่าง ๆในโลก แต่ในที่นี้มโนทัศน์หลักจะกำหนดไว้โดยละเอียดและครอบคลุมเพื่อที่จะให้ทุกภาษาสามารถติดต่อกันได้ด้วยมโนทัศน์หลักนี้

**2.2 ความสัมพันธ์** ความสัมพันธ์แสดงว่ามโนทัศน์สองมโนทัศน์มีความสัมพันธ์กันอย่างไร ความสัมพันธ์ดังกล่าวแบ่งออกเป็นแบบต่าง เช่น การกสัมพันธ์ ความสัมพันธ์ของเหตุการณ์ ความสัมพันธ์ทางความหมาย ความสัมพันธ์เทียม เป็นต้น

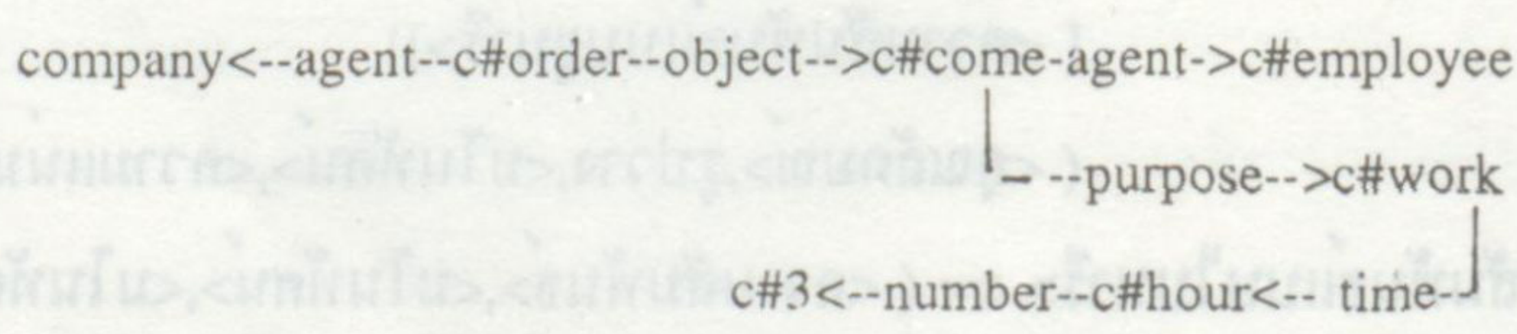
ต่อไปนี้เป็นรายละเอียดของความสัมพันธ์ที่ใช้แสดงความหมายในภาษาไทย

**2.2.1 การกสัมพันธ์** เป็นความสัมพันธ์ที่แสดงการกสัมพันธ์ระหว่างมโนทัศน์ 2 มโนทัศน์

**agent** : ใช้แสดงความสัมพันธ์ว่าเป็นผู้กระทำเหตุการณ์อย่างตั้งใจ ผู้กระทำที่มีความตั้งใจได้แก่ มนุษย์ สัตว์ องค์กร คอมพิวเตอร์ ปรากฏการณ์ธรรมชาติ และสิ่งที่เคลื่อนที่ด้วยเครื่องยนต์

ตัวอย่าง ก. ผู้กระทำที่เป็นองค์กรเราถือว่าองค์กรเป็นนิติบุคคลที่ทำหน้าที่เหมือนบุคคลตามกฎหมาย

ป.ย. 1 บริษัทให้พนักงานมาทำงาน 8 โมง



ข. ผู้กระทำที่เป็นปรากฏการณ์ธรรมชาติ ในที่นี้เราถือว่าปรากฏการณ์ธรรมชาติเกิดขึ้นเองโดยไม่มีผู้ใดจงใจให้เหตุการณ์นั้นเกิด เป็นการเกิดเองตามธรรมชาติ ดังนั้นจึงกล่าวได้ว่าปรากฏการณ์ธรรมชาติทำให้ก่อเหตุการณ์โดยความตั้งใจ

ป.ย. 2 ฝนตกมาก

c#rain <--agent-- c#fall --manner-->c#hard

ค. ผู้กระทำที่เป็นสิ่งที่เคลื่อนที่ได้ด้วยเครื่องยนต์ เช่นรถยนต์ที่ขับเคลื่อนด้วยเครื่องยนต์หรือโปรแกรม

ป.ย. 3 รถไฟวิ่งเร็วมาก

c#train<--agent--c#run--manner--c#fast<--object--c#very

***object*** : แสดงความสัมพันธ์ที่เป็นผู้ถูกกระทำ หรือผู้ทรงสภาพ

ตัวอย่าง ก. ผู้ถูกกระทำหมายถึง ผู้รับผลจากการเปลี่ยนแปลงของเหตุการณ์

ป.ย. 4 เราซื้อหนังสือที่ร้านหนังสือ

c#we<--agent--c#buy--object-->c#book
|--place-->c#bookshop

ข. ผู้ทรงสภาพ หมายถึงการเป็นเงื่อนไขของเหตุการณ์ หรือบอกการคงอยู่ของสิ่งหนึ่ง

ป.ย. 5 มีกระดาษบนโต๊ะ

c#exist--object-->c#paper<--object--c#nil--place-->c#table

ค. ผู้รองรับความสัมพันธ์ บางครั้งเราต้องการแสดงความมาก-น้อยของคุณลักษณ์ เราจะใช้ 'object' เพื่อแสดงความสัมพันธ์แบบนี้

ป.ย. 6 ดอกไม้สวยมาก

c#flower<--a-object--c#beautiful<--object--c#very

***a-object*** : คุณลักษณ์ของวัตถุ ความสัมพันธ์แบบนี้แสดงถึงคุณสมบัติของวัตถุที่หมายรวมถึง รูปร่าง ลักษณะ สี รสชาติ ธรรมชาติ และการประเมินผลรวมทั้งการบรรยาย

ตัวอย่าง ก. รูปร่าง

ป.ย. 7 ลูกบอลกลม

c#ball<--a-object--c#sphere

ข. สี

ป.ย. 8 ลูกบอลแดง

c#ball<--a-object--c#red

ค. การประเมินผล

ป.ย. 9 แม่เป็นคนดี

c#mother<--a-object--c#person--a-object-->c#good

ง. การบรรยาย

ป.ย.10 พ่อเป็นเซลล์แมน

c#father<--a-object--c#salesman

***manner*** : วิธีที่การกระทำหรือปรากฏการณ์เปลี่ยนแปลง ใช้เสดงวิธีการเปลี่ยนแปลง ผลของการกระทำ หรือช่วงเวลาความถี่ของเหตุการณ์

ตัวอย่าง ก. วิธีการของการกระทำ

ป.ย. 11 นักเรียนเรียนอย่างตั้งใจ

c#student<--agent--c#learn--manner-->intentionally

ข. ผลของการกระทำ

ป.ย. 12 นักเรียนเรียนหนักมาก

c#student<--agent--c#learn--manner-->c#hard<--object--c#very

ค. ช่วงเวลาของความถี่ของเหตุการณ์

ป.ย. 13 นักเรียนอยู่ที่ห้อง 3 ชั่วโมง

c#student<--agent--c#stay --manner-->hour---number--c#3

|

place

↓

c#room

***implement***: เครื่องมือหรือวิธีการในการกระทำที่ทำให้เหตุการณ์เปลี่ยนแปลง

ตัวอย่าง ก. เครื่องมือหมายถึงอุปกรณ์ที่ผู้กระทำใช้กระทำเหตุการณ์

ป.ย. 14 เขาขว้างก้อนหินใส่ผู้คน

c#he<--agent--c#throw--implement-->c#stone

goal

c#c#people

ข. วิธีการหมายถึงวิธีการที่ใช้ในการกระทำรวมถึง วิธีการ ขบวนการ หรือสิ่งของที่ใช้ในการกระทำ

ป.ย. 15 ฉันเดินทางด้วยรถไฟ

c#I<--agent--c#travel--implement-->c#train

***material*** :วัสดุที่ใช้เป็นส่วนประกอบหรือองค์ประกอบของสิ่งของหรือสภาพหนึ่ง ๆ ความสัมพันธ์นี้ใช้เพื่อแสดงว่าเป็นส่วนประกอบในการสร้างสิ่งของ หรือใช้เพื่อแสดงองค์ประกอบของสิ่งของ

ตัวอย่าง ก. วัสดุที่ใช้ประกอบการสร้างสิ่งใหม่ ในภาษาไทยส่วนใหญ่จะใช้กริยาวลีว่า "ทำด้วย"

ป.ย.16 เก้าอี้ทำด้วยพลาสติก

c#chair<--object--c#made of--material-->c#plastic

ข. องค์ประกอบของวัตถุหมายถึงวัตถุนั้นประกอบไปด้วยองค์ประกอบอะไร

ป.ย.17 น้ำประกอบด้วยไฮโดรเยนและออกซิเจน

c#compose ---material-->(--main-->c#hydrogen--and-->c#oxygen)

--object -->water

***time*** : เวลาที่เหตุการณ์นั้นเกิด ใช้ในความหมายทั้งเวลาที่เกิดขึ้นจริง และเวลาสัมพัทธ์

ตัวอย่าง ก. เวลาจริง หมายถึงเวลาที่เหตุการณ์นั้นเกิดขึ้นตามเวลาจริง เช่น เวลาที่บ่งได้เป็น วันที่ เดือน ปี ฤดู หรือวันเทศกาล

ป.ย.18 เขาเกิดวันที่ 7 สิงหาคม 1960

main----> <----main

c# born --time-->(c#date--modifier-->7

and

c#month--modifier-->c#august

and

c#year--modifier-->c#1960 )



สวทช.
คอ.8
0109
2536
ฉ.2
รวบ 000715 No.
418.02
ก154
2536
ฉ.2

ข. เวลาสัมพัทธ์ หมายถึงเวลาที่สัมพันธ์กับเหตุการณ์อื่นหรือสถานที่อื่น

ป.ย.19 นักเรียนเล่นกันตอนครูสอนหนังสือ

c#play ---time-->c#teach --agent-->c#teacher

|

--agent -->c#student

***time-from*** : เวลาที่เหตุการณ์เริ่มต้น ใช้เพื่อแสดงเวลาที่เหตุการณ์เริ่มต้น

ป.ย. 20 เขามาถึงตั้งแต่ 5 โมง

c#reach--time-from-->c#o"clock--modifier-->c#5

|

--agent-->c#he

***time-to*** : เวลาที่เหตุการณ์สิ้นสุด ใช้เพื่อบอกความสัมพันธ์ของเวลาที่สิ้นสุด

ป.ย.21 เขาทำการบ้านจนถึงเที่ยงคืน

c#he<--agent--c#do--time-to-->c#midnight

***place*** : สถานที่ทางกายภาพที่เหตุการณ์เกิด สถานที่ทางกายภาพหมายถึงสถานที่ที่เหตุการณ์เกิดขึ้น เช่น ถนน ทุ่งนาและสามารถแสดงได้ว่าวัตถุหนึ่งปรากฏอยู่ในสถานที่นั้น ๆ

ป.ย.22 ก. นักเรียนเรียนอยู่ในมหาวิทยาลัย

c#student<--agent--c#learn--place-->c#university

ป.ย.23 ข. นักเรียนทำงานวิจัยในสาขาวิทยาศาสาตร์

c#student<--agent--c#do --place-->c#field--modifier-->c#science

|

object

↓

c#research work

***scene*** : สถานที่ทางรูปธรรมที่เหตุการณ์นั้นเกิด หมายถึงสถานที่ที่ได้รับการตีความว่ามีเหตุการณ์เกิดขึ้น

ป.ย.24 เธอร้องเพลงทางทีวี

c#she<--agent--c#sing --scene--c#T.V.

ป.ย.25 เธอทำงานทางประชาสัมพันธ์

c#she<--agent--c#work--scene-->c#advertising

***condition*** : เงื่อนไขการเกิดเหตุการณ์ ใช้แสดงความสัมพันธ์แบบเงื่อนไขหรือการก่อ

ให้เกิดเหตุการณ์

ตัวอย่าง ก. เงื่อนไขของเหตุการณ์ (conditional event)

ป.ย.26 คุณต้องทำงานให้เสร็จภายในเดือนนี้

c#do --purpose-->c#complete --condition--> c#month---->this

--agent-->c#you

ข. เหตุการณ์เป็นเหตุเป็นผลกัน (causal event)

ป.ย.27 ฝนตกหนักมากน้ำจึงท่วม

c#flood--object-->c#water

--condition-->c#fall --manner--> c#hard

--agent-->c#rain

***cooccurence*** : การเกิดพร้อมกันแสดงความสัมพันธ์ที่เหตุการณ์ 2 เหตุการณ์เกิดขึ้นพร้อมกัน เช่น

ป.ย.28 ฉันนั่งกินไอศครีม

c#I<--agent--c#sit--cooccurrence-->c#eat--object-->c#icecream

***sequence*** : เวลาที่ต่อเนื่อง หมายถึงการแสดงเหตุการณ์ที่เกิดขึ้นต่อเนื่องกัน เช่น

ป.ย.29 ฉันนั่งลงแล้วก็วาดการ์ตูน

c#I<--agent--c#sit down <--sequence--c#draw--object-->c#cartoon

***quantity*** : หมายถึงปริมาณของจำนวนสิ่งของที่เปลี่ยนไปในเหตุการณ์หรือจำนวนปริมาณของปรากฏการณ์ บางอย่าง

ตัวอย่าง ก. ปริมาณสิ่งของ

ป.ย. 30 ส้ม 1 กิโล

c#orange --quantity-->c#kilogram--number-->c#1

ข. ปริมาณของการเปลี่ยนแปลง

ป.ย.31 ฉันน้ำหนักลดลง 3 กิโล

c#I<--agent--c#lose--quantity-->c#kilogram--number-->c#3

***number*** : จำนวนนับของสิ่งของ ใช้เพื่อแสดงปริมาณของสิ่งของด้วยจำนวนนับ

ป.ย.32 นกบินมา 3 ตัว

c#fly--agent-->c#bird--number-->c#3

***basis*** : ฐานของความสัมพันธ์ ใช้เพื่อแสดงการเปรียบเทียบ

ป.ย.33 บานฉันสวยกว่าบานเธอ

c#beautiful<--object--c#more--basis-->c#house---possesor-->c#you

***and*** : การเชื่อมมโนทัศน์หรือเหตุการณ์ในทางคล้อยตามกัน

ตัวอย่าง ก. การเชื่อมเหตุการณ์ 2 เหตุการณ์

ป.ย.34 วันนี้อากาศเย็นและลมแรง

(--main-->c#cloudy-- and-->c#windy)--a-object-->c#weather

ข. การเชื่อมมโนทัศน์ 2 มโนทัศน์

ป.ย.35 ฉันซื้อหนังสือและสมุดมากมาย

(main---->c#book--and-->c#notebook)<--a-object--c#many

***or*** : การเชื่อมแบบเลือกอย่างใดอย่างหนึ่ง ใช้แสดงเพื่อการเลือกอันเนื่องมาจากความต่างกันของมโนทัศน์ 2 มโนทัศน์ หรือเหตุการณ์ 2 เหตุการณ์

ตัวอย่าง ก. การเลือกระหว่างมโนทัศน์ 2 มโนทัศน์

ป.ย. 36 คุณ ชอบ นก หรือ แมว

c#like--object-->(main---->c#bird--or-->c#cat)

--agent-->c#you

ข. การเลือกระหว่างเหตุการณ์ 2 เหตุการณ์

ป.ย.37 คุณ จะ ไป หรือ ไม่ ไป

main---->(c#go--or-->c#go---->not)--agent-->c#you

***purpose*** : จุดประสงค์ของการกระทำ หมายถึงการแสดงถึงการกระทำที่เป็นจุดประสงค์ของผู้ทำว่ามุ่งที่จะทำ หรือ ไม่ทำเหตุการณ์นั้น

ป.ย.38 พ่อต้องทำงานหนัก เพื่อเลี้ยงลูก

c#work---purpose-->c#take care--object-->c#children

agent -->c#father

***modifier:*** การขยายมโนทัศน์ ใช้เพื่อแสดงการที่มโนทัศน์มีส่วนที่ขยายเป็นอีกมโนทัศน์หนึ่ง

ก. การขยายมโนทัศน์

ป.ย.39 ความก้าวหน้าของชาติ

c#progress---modifier-->c#nation

ข. การแสดงประเภทเดียวกันของมโนทัศน์

ป.ย.40 มะม่วงเป็นผลไม้

c#fruit--modifier-->c#mango

ค. การขยายด้วยำนวนนับ

ป.ย.41 วันที่ 5

c#day--modifier-->c#5

**2.2.2 Pseudo relation** หมายถึงความสัมพันธ์ที่ไม่ได้เกิดขึ้นจริง แต่เป็นการตีความ มีอยู่ 3 แบบดังนี้

*possessor* : การเป็นเจ้าของ ใช้เพื่อแสดงความสัมพันธ์ระหว่างเจ้าของและสิ่งของ

ป.ย.42 หนังสือของฉัน

c#book---possessor-->c#I

ป.ย.43 โทรทัศน์ช่อง 5 เป็นของทหาร

c#T.V.---possesor-->c#military

---modifier-->c#chanel---modifier-->c#5

***from to*** : ขอบเขตของวัตถุและเหตุการณ์ แสดงถึงขอบเขตของเวลา และระยะทางของมโนทัศน์ 2 มโนทัศน์

ตัวอย่าง ก. ขอบเขตของเวลา

ป.ย.44 เราทำงานตั้งแต่ 5 โมงถึง 7 โมง

c#work --manner-->(main---->c#o'clock--from-to-->c#o'clock

--agent-->c#we --modifier-->c#7 --modifier-->c#5)

ข. ช่วงของระยะทาง

ป.ย.45 เราเดินทางจากกรุงเทพถึงโตเกียว

c#travel--modifier-->(main--->c#tokyo---from-to-->c#bangkok)

--agent-->c#we

**unit** : หน่วยของสิ่งของและเหตุการณ์ ใช้เพื่อแสดงเกณฑ์มาตรฐานของการวัด เช่น เมื่อเราต้องการบอกมาตรฐานของสิ่งของตามหน่วยมาตราวัด

ป.ย.46 เราซื้อมะม่วงกิโลละ 20 บาท

(main---->c#baht---number-->c#20)<--unit--c#kilokram--number-->c#1

c#buy--manner-->

***beneficiary*** : บุคคลที่ได้รับประโยชน์จากการกระทำใช้เพื่อแสดงความสัมพันธ์ของบุคคลที่ได้รับประโยชน์หรือเสียผลประโยชน์จากการกระทำอย่างใดอย่างหนึ่ง

ป.ย.47 พ่อทำงานเพื่อลูก

c#work---beneficiary-->c#children

**2.2.3 Semantic relation** เป็นความสัมพันธ์ทางความหมายของมโนทัศน์ ได้แก่

part-of : เป็นการแสดงเรื่องความสัมพันธ์แบบส่วนใหญ่ - ส่วนย่อย โดยทั่วไปความสัมพันธ์แบบนี้จะปรากฏอยู่ในพจนานุกรมประเภท concept classification แต่ถ้าพบความสัมพันธ์นี้ปรากฏอยู่ในโครงสร้างประโยคให้ใช้"modifier"แทน

ตัวอย่าง ก. ความสัมพันธ์ทางความหมาย

ป.ย.48 หลังคาบ้านผุแล้ว

c#roof--part of-->c#house

ข. ในโครงข่ายภาษากลางที่ปรากฏ
หลังคาบ้านผุแล้ว

c#house---modifier-->roof

*kind-of* : เป็นความสัมพันธ์แบบชนิดหนึ่งของ บอกความสัมพันธ์ที่เป็น upper class และ lower class ใช้อธิบายความสัมพันธ์ทาง paradigmatic ในเรื่องลำดับของมโนทัศน์เพื่อบอกการแบ่งประเภทมโนทัศน์ ความสัมพันธ์ชนิดนี้จะเก็บไว้ในพจนานุกรม

เช่น c#human--kind of-->c#animal

**2.3. คุณลักษณ์ (Attributes)** คุณลักษณ์แสดงถึงคุณสมบัติของมโนทัศน์เกี่ยวกับการบอกระยะใกล้ ไกล เวลา การณ์ลักษณะ การปฏิเสธ ความเห็นและเจตนาของผู้พูด การบอกจำนวน เป็นต้น

รูปแบบของการแสดงคุณลักษณ์ มี 2 รูปแบบคือเป็นความสัมพันธ์ที่ชี้ออก หรือชี้เข้ามโนทัศน์ดังนี้

ก. มโนทัศน์---คุณลักษณ์-->

ข. หมายถึง<---คุณลักษณ์--

**2.3.1 คุณลักษณ์ที่แสดงการบอกระยะใกล้ ไกล** ความสัมพันธ์ที่ใช้มี 2 ความสัมพันธ์คือ

***that:*** บอกระยะไกล

ป.ย.49 นกตัวโน้นสวย

c#beautiful--a-object-->c#bird---->that

***this*** : บอกระยะใกล้

ป.ย.50 แมวตัวนี้กินปลาไป

c#eat--agent-->c#cat---->this

**2.3.2 คุณลักษณ์ที่บอกจำนวนที่กำหนดไม่ได้** ความสัมพันธ์ที่ใช้มี 3 ความสัมพันธ์คือ

***all*** : แทนทุกหน่วยที่เป็นของมโนทัศน์

ป.ย.51 แมวทุกตัวชอบกินปลา

c#like--agent-->c#cat---->all

*generic* : แสดงความทั่วไปของมโนทัศน์

ป.ย.52 เขาชอบแอปเปิ้ล

c#he<--agent--c#like--object-->c#apple---->generic

**2.3.3 คุณลักษณ์ที่แสดงการปฏิเสธ** มีใช้อยู่ 1 แบบคือ

***not*** ใช้บอกการปฏิเสธมโนทัศน์ หรือ ปฏิเสธเหตุการณ์ หรือปฏิเสธคุณลักษณ์

ป.ย.53 เขาไม่ชอบหมอฟัน

c#he<--agent--c#like--object-->c#dentist

------>not

**2.3.4 คุณลักษณ์ใช้แสดง กาล** คุณลักษณ์ที่กำหนดขึ้นแทนกาล แบ่งตามมุมมองของผู้พูดเป็น 3 คุณลักษณ์ด้วยกันคือ

past : แสดงเหตุการณ์ว่าเกิดขึ้นในเวลาที่ผ่านมา

present : แสดงเหตุการณ์ว่าเกิดขึ้นในเวลาปัจจุบัน

future : แสดงเหตุการณ์ว่าเกิดขึ้นในเวลาที่ยังมาไม่ถึง

2.3.5 คุณลักษณ์ที่แสดงการณ์ลักษณะ (aspect)

| รูปแทนการณ์ลักษณะของภาคแสดง | ตัวอย่างภาษาไทย |
|---|---|
| - การณ์ลักษณะบอกเหตุการณ์ที่ยังไม่เริ่มต้น<br>(c#pred--begin-->)--yet--> | ยังไม่ได้ กริยา |
| - การณ์ลักษณะบอกเหตุการณ์ก่อนการเริ่มต้น<br>(c#pred--begin-->)--soon--> | กำลังจะ กริยา<br>เกือบจะ กริยา |

กำลังเริ่มจะ กริยา

กำลังจะเริ่ม กริยา

- การณ์ลักษณะบอกเหตุการณ์การเริ่มต้น เริ่ม กริยา

c#pred--begin-->

- การณ์ลักษณะบอกเหตุการณ์หลังเริ่มต้น เพิ่ง กริยา

(c#pred --begin-->)--just--> เพิ่งเริ่ม กริยา

เพิ่งจะ กริยา

- การณ์ลักษณะบอกเหตุการณ์ที่กำลังดำเนินอยู่ กำลัง กริยา

c#pred --progress--> กำลัง กริยา อยู่

- การณ์ลักษณะบอกเหตุการณ์ที่ต่อเนื่อง ยัง กริยา อยู่

c#pred --continue--> ทุกวัน

- การณ์ลักษณะบอกเหตุการณ์ก่อนจะเสร็จสิ้น กริยา เกือบ จะ เสร็จ

(c#pred –end-->) --soon--> กริยา เกือบ เสร็จ แล้ว

- การณ์ลักษณะบอกเหตุการณ์เสร็จสิ้น กริยา เสร็จ แล้ว

(c#pred –end-->) –already-->

- การณ์ลักษณะบอกเหตุการณ์สมบูรณ์ กริยา จบ

c#pred –complete--> กริยา ครบ

กริยา สำเร็จ

- การณ์ลักษณะบอกเหตุการณ์เพิ่งจะเสร็จสิ้น เพิ่ง กริยา เสร็จ

(c#pred –end-->) –just-->

| | |
|---|---|
| - การณ์ลักษณะบอกเหตุการณ์ที่บอกสภาพ | กริยา อยู่ |
| c#pred --state--> | กริยา ไว้ |
| - การณ์ลักษณะบอกเหตุการณ์ที่ยังไม่เสร็จ | ยัง กริยา ไม่ เสร็จ |
| (c#pred --end-->)--yet--> | กริยา ยัง ไม่ เสร็จ |
| - การณ์ลักษณะบอกเหตุการณ์ยังไม่สิ้นสุด | กริยา ไม่ จบ |
| (c#pred –complete->) –yet-> | กริยา ไม่ ครบ |
| | กริยา ไม่ สำเร็จ |

**2.3.6 คุณลักษณ์ที่แสดงความตั้งใจและเจตนาของผู้พูด** คุณลักษณ์ดังกล่าวเป็นคุณลักษณ์ที่ขยายประโยคทั้งประโยค ดังนั้นรูปแทนของประโยคที่ประกอบไปด้วยคุณลักษณ์แสดงเจตนาจะแสดงดังนี้

ตัวอย่าง

ป.ย. 54 ผมต้องจอดรถไว้ข้างล่าง

ต้อง แสดงคุณลักษณ์ "duty"

state----> --place-->c#downstair

c#I<--agent--c#park--object-->c#car

st

c#statement ---->duty

**ตารางคุณลักษณ์แสดงความตั้งใจและเจตนาของผู้พูด**

| คุณลักษณ์ | รูป | ประเภทของคำ | ตัวอย่าง |
|---|---|---|---|
| 1. OBLIGATION | | | |
| -IMPERATIVE | จง | AUX | จงกากบาทหน้าข้อที่ถูก |

| | | | |
|---|---|---|---|
| | ห้าม | AUX | ห้ามสูบบุหรี่ |
| -REQUEST | กรุณา | AUX | กรุณาทำให้ฉันด้วย |
| | ช่วย | AUX | ช่วยวักให้ที |
| -INVITE | กันเถอะ | ADV | เราไปกินข้าวกันเถอะ |
| | มา (กริยา)กัน | AUX | มาร้องเพลงกัน |
| -DUTY | ต้อง | AUX | ฉันต้องจอดรถไว้ข้างล่าง |
| | จำเป็นต้อง | AUX | นักกายกรรมจำเป็นต้องรู้เรื่อง วิทยาศาสตร์ |
| -SHOULD | ควร | AUX | มันควรเป็นคืนแห่งการฉลอง |
| -ADVICE/ -RECOMMEND | น่า | AUX | สิ่งนี้น่าเป็นตัวอย่างที่ดี |
| 2. THREAT | | | |
| -NATURAL - THING | นิยม | AUX | คนไทยนิยมกินหมาก |
| | มัก | AUX | คนแก่มักไปวัดในวันพระ |
| | ชอบ | AUX | คนญี่ปุ่นชอบกินปลาดิบ |
| 3. GRANT | | | |
| -GRANT | ยอมให้ | V | เขายอมให้เธอไป |
| | อนุญาตให้ | V | ผู้จัดการอนุญาตให้ลูกน้องลาได้ |
| -CONSENT | ยินดี | V | คุณยินดีที่จะไปไหม |
| -SUFFICIENCY | ก็พอ | ADV | ไปเยี่ยมก็พอ |
| | เพียง | ADV | เพียงโทรศัพท์มาก็พอ |

| | | | |
|---|---|---|---|
| -GRANT-NOT | ไม่ยอมให้ | V | คุณไม่ยอมให้เธอไป |
| | ไม่อนุญาตให้ | V | คุณไม่อนุญาตให้เธอไป |

4. AGREEMENT

| | | | |
|---|---|---|---|
| -NATURAL- | จึง | AUX | ฝนตกน้ำจึงท่วม |
| RESULT | เพราะ | CONJ | น้ำท่วมเพราะฝนตก |
| | ก็เลย | AUX | ฝนตกน้ำก็เลยท่วม |

5. REQUIRE-AGREEMENT

| | | | |
|---|---|---|---|
| - REQUIRE- | จริงไหม | EITT | ที่นี่หนาวจริงไหม |
| AGREEMENT | ใช่ไหม | EITT | ที่นี่หนาวใช่ไหม |

6. JUDGEMENT

| | | | |
|---|---|---|---|
| -IF | หาก | CONJ | หากเขาไม่มา ฉันต้องตาย |
| -THOUGHT | ถ้า | CONJ | ถ้าฉันเป็นนก ฉันจะบินไปให้ไกล |

7. PROBABILITY

| | | | |
|---|---|---|---|
| -MAYBE | บางที | ADV | บางทีเขาคงมา |
| | อาจ | AUX | เขาอาจไม่รู้ |
| -SEEM | คาดว่า | V | คาดว่าพรุ่งนี้ฝนจะตก |
| -SURE | แน่ | ADV | เขามาแน่ |
| - RUMOR | ลือกันว่า | V | ลือกันว่าเธอสวยมาก |
| | เล่ากันว่า | V | เล่ากันว่าฟูจีสวยมาก |

- APPEARANCE

| | | | |
|---|---|---|---|
| | ทำท่า | V | ฝนทำท่าจะตก |
| -CONCLUDE | | | |
| | จึง | V | นักเรียนไม่ดีครูจึงตี |
| 9.FEELING | | | |
| -BLAME | ไม่น่า | AUX | เขาไม่น่าทำเลน |
| | ไม่ควร | AUX | เขาไม่ควรทำแบบนี้ |
| -EXCLAMATION | | | |
| | อะไรอย่างนี้ | | สวย อะไรอย่างนี้ |
| -PITY | น่าสงสารที่ | | น่าสงสารที่เขาไม่ได่ไป |
| -REGRET | ถึงแม้ว่า | CONJ | ถึงแม้ว่าเขาจะรวย เขาไม่มีความสุข |
| -UNDER-ESTIMATE | แค่ | ADV | แค่เดินไปเท่านั้น |
| -UNEXPECTED | | | |
| | ไม่นึกไม่ฝัน | V | ไม่นึกไม่ฝันว่าจะเดินได้ |
| 10. FEELINGS FOR BENEFIT | | | |
| -GIVE-BENEFIT | | | |
| | ให้ | AUX | เขาทำการบ้านให้ |
| | ช่วย | AUX | เขาช่วยเธอทำงาน |
| -GET-BENEFIT | | | |
| | ถูก | AUX | หล่อนถูกใช้ให้ไปซื้อของ |
| 11. INTERROGATION | | | |
| -INTERROGATION | | | |
| | ทำไม | PRON | ทำไมเขาไม่มา |

12. RESPECT

-RESPECT

ครับ ค่ะ PARTICLE จะกลับแล้วหรือครับ

-POLITE

โปรด AUX โปรดถอดรองเท้า

**4. การแทนรูปประโยคด้วยภาษากลาง** โครงสร้างภาษากลางแสดงโครงสร้างทางความหมายของประโยคได้โดยใช้มโนทัศน์ c#statement เชื่อมกับกริยาเพื่อบอกความเป็นประโยค ทั้งยังเป็นที่เก็บข้อมูลที่ขยายความทั้งประโยคได้ นอกจากนั้นยังมีความสัมพันธ์ที่แสดงการเชื่อมต่อระหว่างประโยค เช่น การเป็น sub statement หรือ co-statement ของ ประโยค 2 ประโยคเช่น

1. การแสดง statement

ป.ย.55 บ้าน นี้ สวย

<--this--c#house<---a-object--c#beautiful

st

c#statement

ข. การแสดงการเชื่อมประโยคด้วย sub-statement

ป.ย. 56 เขาไปทำงานที่บริษัท

main----> ---place-->company

he <---agent--c#move---purpose-->work

go<----

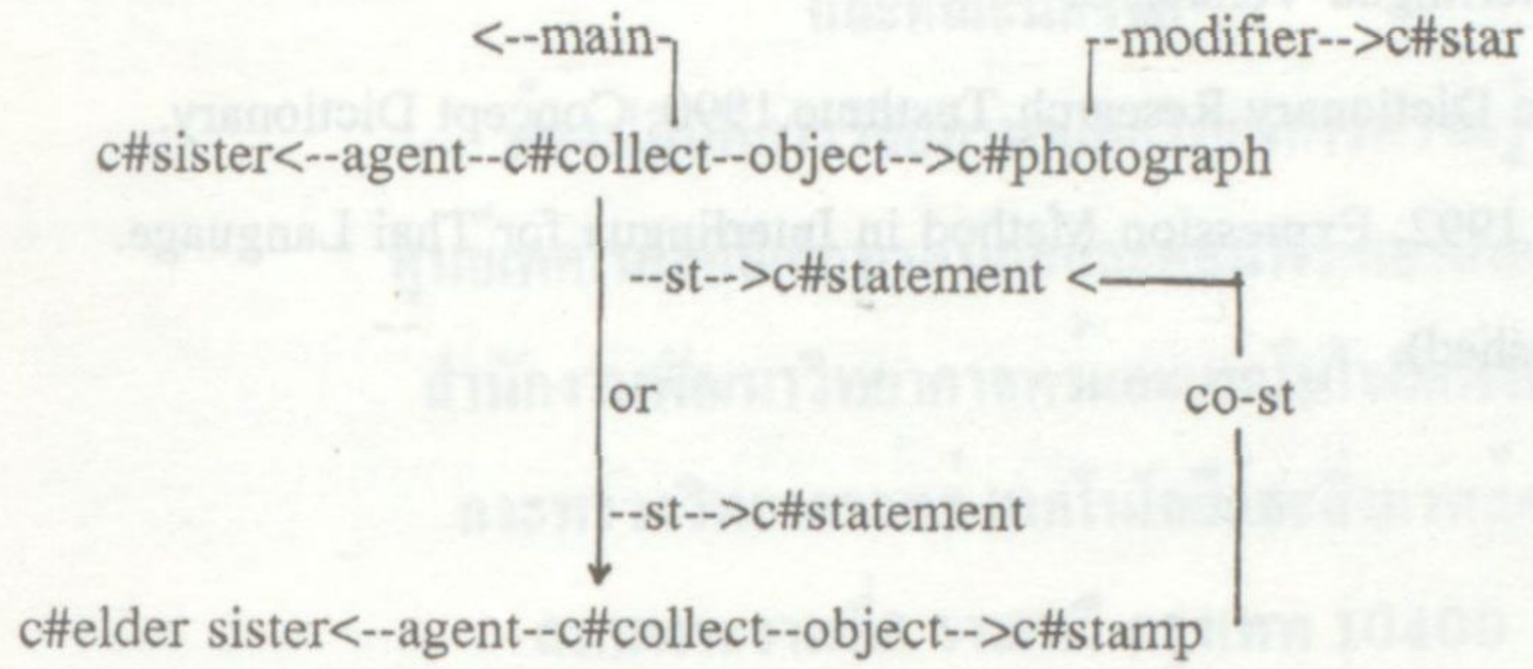
st st

c#statement c#statement

---sub-st---->

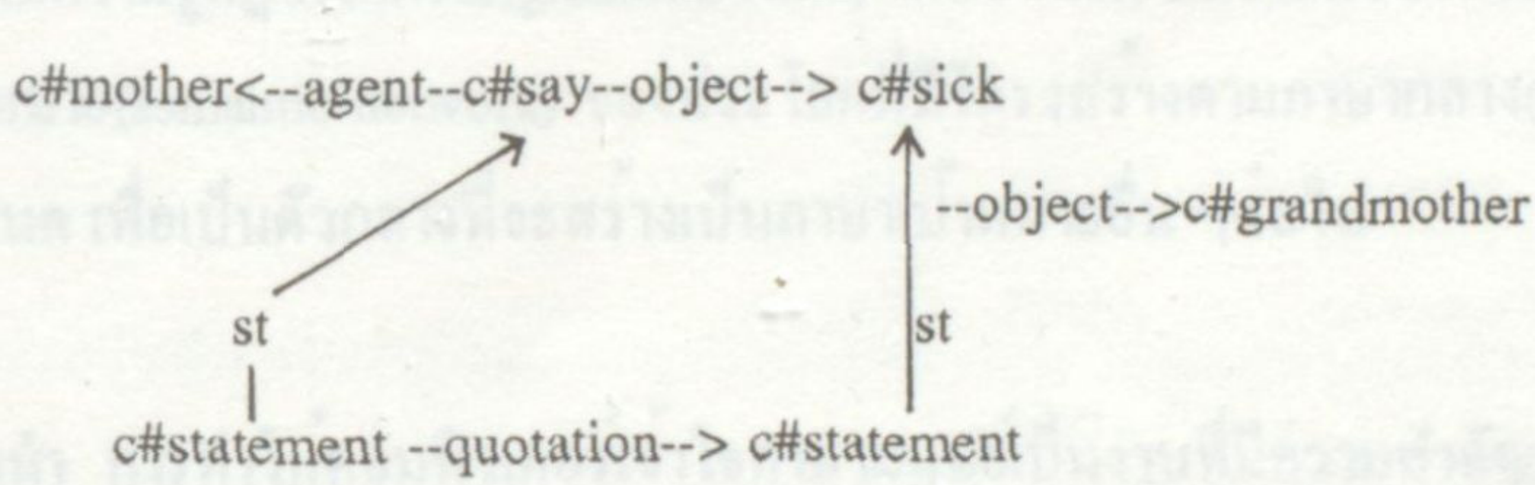
ค. การแสดงการเชื่อมประโยคด้วย co-statement

ป.ย. 57 น้องสาวสะสมรูปภาพดารา แต่พี่สาวสะสมแสตมป์




ง. การแสดงการเชื่อมประโยคด้วย 'quotation' กรณีที่ประโยคหนึ่งเป็น คำกล่าวของอีกประโยคหนึ่ง เช่น

ป.ย. 58 แม่พูดว่า "คุณยายไม่สบาย"




**4. สรุป** การแสดงรูปแทนภาษากลางนี้เป็นรูปแทนที่ได้กำหนดขึ้นจากการค้นคว้าของนักวิจัยที่ EDR ประเทศญี่ปุ่น และพัฒนาร่วมกันโดยนักวิจัย 5 ชาติที่เข้าร่วมโครงการเครื่องแปลภาษา ภาษากลางที่เสนอในที่นี้เป็นภาษากลาง version 3 (Uchida :1991) อย่างไรก็ตาม ภาษากลางยังจะมีการปรับปรุงกันโดยนักวิจัย 5 ชาติเพื่อที่จะหาวิธีการแทนความหมายที่เป็นสากลมากยิ่งขึ้น

## บรรณานุกรม

CICC, 1992, Interlingua Version 3.

Japan Electronic Dictionary Research Tnstitute,1990. Concept Dictionary.

Phantachat, W, 1992, Expression Method in Interlingua for Thai Language. (unpublished).

# ระบบการวิเคราะห์ภาษาไทยเพื่อการแปลภาษาด้วยคอมพิวเตอร์

**วันทนีย์ พันธชาติ**

**และคณะนักวิจัย***

**ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้**

**ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ**

**สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ**

**กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม**

**ถนนพระรามที่6 ราชเทวี กรุงเทพ 10400**

**บทคัดย่อ** ระบบการวิเคราะห์ภาษาไทยนี้เป็นส่วนหนึ่งของระบบเครื่องแปลภาษาแบบพหุภาษา(Multi-lingua Machine Translation) ทำหน้าที่วิเคราะห์ประโยคอินพุตภาษาไทยโดยใช้ฐานความรู้(data base) สองส่วนคือฐานความรู้พจนานุกรม(dictionary) และฐานความรู้กฎไวยากรณ์(grammar rules) ผลจากการวิเคราะห์จะได้โครงข่ายทางความหมาย(semantic network) ของประโยคที่มีโครงสร้างตามภาษากลาง(Interlingua) ที่กำหนด เพื่อเป็นตัวกลางที่จะสร้างเป็นภาษาเป้าหมายอื่น ๆต่อไป

**1. บทนำ** การทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์เป็นงานที่มีความสำคัญประการหนึ่งในยุคของเทคโนโลยีสารสนเทศนี้ เพราะจะช่วยให้มนุษย์สามารถสื่อสารหรือโต้ตอบกับเครื่องคอมพิวเตอร์ได้ และสามารถนำคอมพิวเตอร์มาใช้ในงานด้านการแปลภาษาอันเป็นความใฝ่ฝันสูงสุดของนักวิชาการในปัจจุบัน การพัฒนาระบบการวิเคราะห์ภาษาเป็นส่วนหนึ่งของความพยายามที่จะทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์

**1.1 เป้าหมายของการวิเคราะห์ภาษา** การวิเคราะห์ภาษามีเป้าหมายเพื่อวิเคราะห์และจดจำความหมาย(recognize) ของภาษาอินพุต รูปแทนที่ใช้แสดงความหมายของประโยคอินพุตทางคอมพิวเตอร์มีหลายระบบ ระบบที่คณะผู้เขียนนำมาใช้ในที่นี้เป็นระบบหนึ่งที่เรียกว่า ภาษากลาง(interlingua) ซึ่งเหมาะสำหรับระบบการแปลแบบพหุภาษา ในปัจจุบันการกำหนดภาษากลางเพื่อใช้เป็นรูปแทนทางความหมายนี้อยู่

**2.1 ไวยากรณ์โครงสร้างวลี** แนวคิดสำคัญของไวยากรณ์โครงสร้างวลีนี้คือ การรวมหน่วยย่อยให้เป็นหน่วยที่ใหญ่กว่า ซึ่งก็คือ การรวมคำให้เป็นวลี (phrase) เช่น กริยาวลี (verb phrase) และนามวลี (noun phrase) เป็นต้น และวลีต่าง ๆจะรวมเป็นหน่วยที่ใหญ่กว่าขึ้นไป ซึ่งก็คือ ประโยค(sentence)

ในการแจงส่วนประโยค ได้อาศัยแนวคิดของทฤษฎีนี้มาสร้างกฎโครงสร้างวลี (phrase structure rules) ขึ้นเพื่อรวมหน่วยต่างๆ (lexical items) ให้มีโครงสร้างเป็นวลีและเป็นประโยค ตัวอย่างของกฎโครงสร้างวลีสำหรับภาษาไทยตามที่คณะผู้วิจัยได้วิเคราะห์ไว้ มีรายละเอียดดังนี้

**กฎโครงสร้างวลีภาษาไทย**

ประโยคความเดียว

S ======> (NP) VP (NP) (NP) (PP) (ADVP)

VP ======> (PRE-AUXP) (NEG) V (POST-AUXP)

NP ======> NP1 PP

NP ======> NP1 CONJ NP1

PP ======> PREP NP1

ADVP ======> ADVI (NALN)

ADVS

ADVN

NP1 ======> N (CLASS) (VATT) (DETP)

PRON

DETP ======> (QDETP) (DEFP)

QDETP======> (DDBQ) DCNM CLASS (DDAQ)

(DIBQ) (DIAQ)

DEFP ======> (DDAN)(CLASS)(DONM)(DDAN)

(ดูชื่อย่อประเภทของคำ (category) ที่ใช้ในการวิเคราะห์ประกอบท้ายบทความ)

ในระหว่างการวิจัยระหว่าง 5 ประเทศที่ร่วมพัฒนาเครื่องแปลภาษาสำหรับภาษาในเอเชีย 5 ภาษา อันได้แก่ประเทศญี่ปุ่น จีน อินโดนีเซีย มาเลเซีย และ ไทย

**1.2 องค์ประกอบของระบบการวิเคราะห์ภาษาไทย** ระบบการวิเคราะห์ภาษาที่ใช้ในปัจจุบันแบ่งเป็น 2 ส่วนคือ

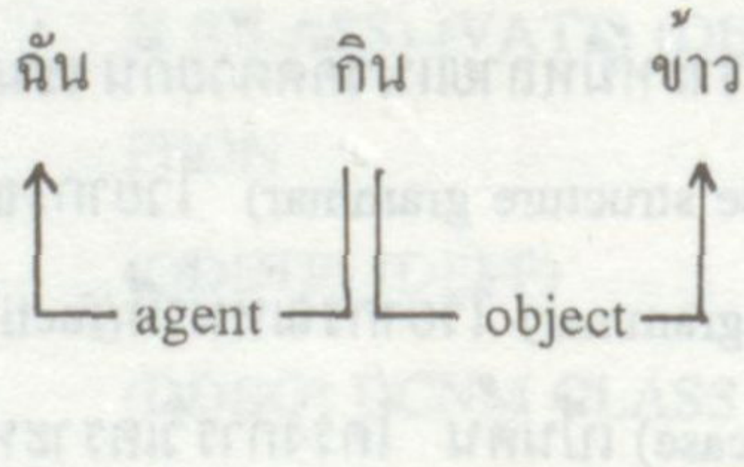
ก. ส่วนการวิเคราะห์คำ (morphology analysis) เนื่องจากภาษาไทยเป็นภาษาคำติดต่อ(agglutinative language) จึงไม่มีการแสดงขอบเขตของคำ ดังนั้นจึงต้องมีส่วนที่จัดการกับการแบ่งคำโดยสร้างโปรแกรมตัดคำ(word segmentation program) โดยอิงรูป(form) ของคำที่ใช้แทนมโนทัศน์(concept) ซึ่งอาจเป็นได้ทั้งคำเดี่ยว และคำประสม และจัดเก็บไว้ในพจนานุกรม

ข. ส่วนการวิเคราะห์เพื่อแจงส่วนทางวากยสัมพันธ์-อรรถศาสตร์(syntactic-semantic parsing) เป็นส่วนที่จัดการเรื่องการแจงส่วน(parse) ประเภทของคำ การวิเคราะห์(analyse) องค์ประกอบทางวากยสัมพันธ์ และการสร้างความสัมพันธ์ (synthesis) ของมโนทัศน์ในประโยคตามความหมายเพื่อให้ได้รูปแทนทางความหมายที่กำหนด

**2. แนวทฤษฎีทางไวยากรณ์และเทคนิคการวิเคราะห์** คอมพิวเตอร์มิได้มีความฉลาดเช่นมนุษย์ การที่คอมพิวเตอร์จะสามารถวิเคราะห์ความหมายของภาษาได้ ก็จะต้องนำความรู้เกี่ยวกับการวิเคราะห์ภาษาที่มีอยู่ในสมองของคนเราใส่ให้ ปัจจุบันความรู้ที่ใช้ในการวิเคราะห์มีหลายแนวคิดด้วยกัน เช่น แนวคิดเกี่ยวกับไวยากรณ์โครงสร้างวลี(phrase structure grammar) ไวยากรณ์การก (case grammar) ไวยากรณ์พึ่งพา(dependency grammar) ไวยากรณ์หน้าที่(fuctional grammar) และไวยากรณ์ศัพท์การก(lexicase) เป็นต้น โครงการวิเคราะห์ภาษาไทยที่คณะผู้เขียนดำเนินการวิจัยอยู่นี้ ใช้ไวยากรณ์โครงสร้างวลี ไวยากรณ์การก และไวยากรณ์พึ่งพา ประกอบกันเพื่อวิเคราะห์ความหมายของประโยค โดยใช้การก(case) ที่กำหนดขึ้นในภาษากลางแสดงความสัมพันธ์ของคำ ส่วนตัวระบบอาศัยเทคนิคและซอฟท์แวร์การวิเคราะห์ของระบบ PIVOT ซึ่งได้รับการสนับสนุนจากบริษัท NEC ประเทศญี่ปุ่น

อย่างไรก็ตาม การแสดงองค์ประกอบของประโยคเป็นวลีต่างๆข้างต้นยังไม่เพียงพอเพราะไวยากรณ์นี้เพียงแต่รวมหน่วยต่างๆขึ้นเป็นวลีและประโยคเท่านั้น ไม่ได้แสดงข้อมูลของโครงสร้างทางภาษาทั้งในแง่ของวากยสัมพันธ์และอรรถศาสตร์ ดังนั้น จึงได้นำเอาแนวทฤษฎีของไวยากรณ์การกเข้ามาช่วยในการแสดงความสัมพันธ์ของวลีต่างๆ

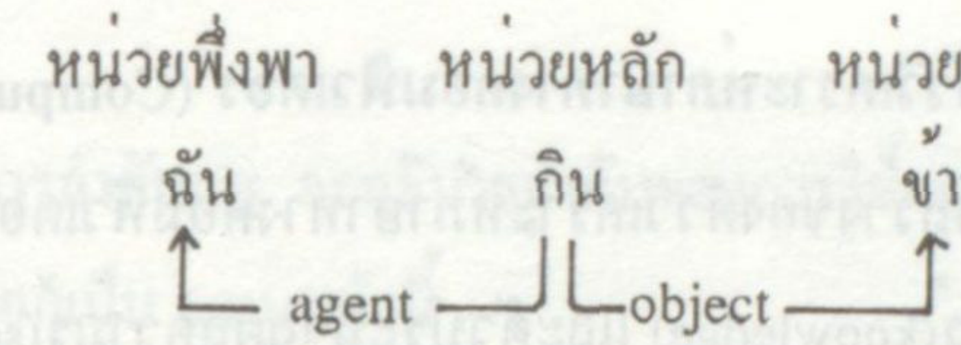
**2.2 ไวยากรณ์การก** แนวคิดสำคัญของไวยากรณ์การกก็คือคำนามในประโยคจะมีการกสัมพันธ์ (case relation) อันเป็นความสัมพันธ์ทางความหมายกับคำกริยาหลักเสมอ และไม่ว่าคำนั้นจะปรากฏในตำแหน่งใดในประโยค ก็ย่อมจะมีการกสัมพันธ์ (เดียวกัน) กับกริยาทั้งสิ้น แนวคิดนี้เป็นแนวคิดสากลของภาษา(universal language) ซึ่งสอดคล้องกับระบบการแปลแบบพหุภาษา การวิเคราะห์ภาษาไทยอาศัยแนวคิดนี้เป็นหลักในการสร้างความสัมพันธ์ทางความหมายของคำกริยาและคำนาม การแสดงการกสัมพันธ์นั้นจะเริ่มจากการหากรอบการก(case frame)ให้กับคำกริยาต่างๆ กรอบการกนี้จะแสดงความสัมพันธ์ระหว่างคำกริยาและคำนามที่อยู่ในรูปแบบบังคับเกิด (obligatory case) ส่วนรูปแบบการกให้เลือกเกิด(optional case) นั้นจะปรากฏในกฎการเลือกการกที่ได้จากการวิเคราะห์ความหมายของคำต่าง ๆที่สัมพันธ์กัน ระบบการแปลภาษาหลายระบบจึงนำแนวคิดนี้มาประยุกต์ใช้ในการหาความสัมพันธ์ทางความหมายของคำกริยาและคำนาม




รูปที่ 1 ตัวอย่างของการกสัมพันธ์ระหว่างคำกริยากับคำนาม

**2.3 ไวยากรณ์พึ่งพา** ไวยากรณ์พึ่งพามีแนวคิดสำคัญคือ หน่วยต่างๆ ย่อมมีความสัมพันธ์ที่พึ่งพากัน(dependency relation ) กล่าวคือ ความสัมพันธ์จะอยู่ในรูปของหน่วยหนึ่งที่ทำหน้าที่เป็นหน่วยหลัก (head) และอีกหน่วยหนึ่งจะเป็นหน่วยขยายหรือ

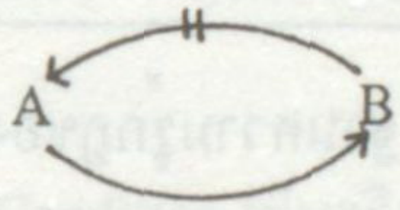
หน่วยพึ่งพา(modifier or depender) แนวคิดการแสดงความสัมพันธ์แบบพึ่งพานี้สามารถให้คำอธิบายทางภาษาได้ใกล้เคียงภาษาธรรมชาติมาก เพราะสามารถแสดงการเชื่อมต่อกันระหว่างหน่วยต่าง ๆ เป็นลำดับ(heirachical link) ซึ่งสอดคล้องกับลักษณะทางวากยสัมพันธ์(syntactic structure) ของภาษาต่าง ๆ เช่น ข้อตกลง (agreement) ระหวางคำกริยากับคำนาม การใชตัวบงชี(marker) และการเรียงลำดับของคำ(word order) ผลของการวิเคราะห์โดยใช้ไวยากรณ์พึ่งพานี้ จะได้โครงข่ายทางความหมายของประโยค ดังตัวอย่าง




รูปที่ 2 ตัวอย่างของความสัมพันธ์แบบหน่วยหลัก-หน่วยพึ่งพา

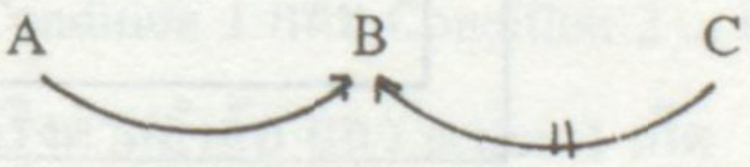
ข้อตกลงเบื้องต้นของการแสดงความสัมพันธ์แบบพึ่งพามีดังต่อไปนี้

ก. ความสัมพันธ์แบบพึ่งพาย่อมไม่เป็นแบบที่ทั้งสองโนด(node)ต่างก็เป็นหน่วยหลัก(head) และหน่วยพึ่งพา(depender) ของกันและกัน(interdependency) เช่น




คำอธิบาย : ในที่นี้ B เป็นหน่วยพึ่งพาของ A ดังนั้น B จะเป็นหน่วยหลักของ A ไม่ได้

ข. หน่วยใด ๆ จะเป็นหน่วยพึ่งพาของหน่วยหลักมากกว่าหนึ่งหน่วยไม่ได้ (double dependency) เช่น

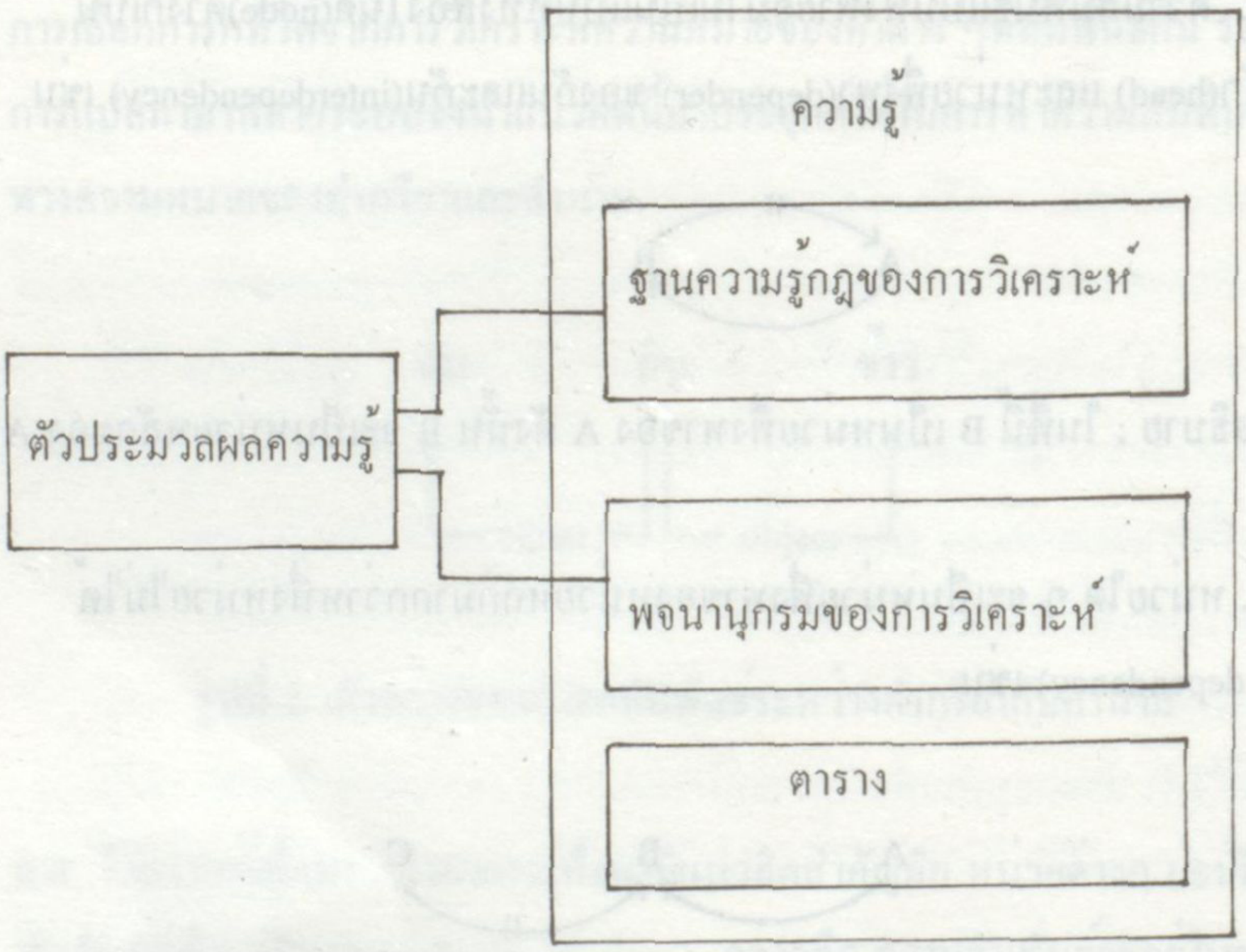



คำอธิบาย : B จะเป็นหน่วยพึ่งพาของทั้ง A และ C ไม่ได้

ค. ทุก ๆ โนดย่อมมีความสัมพันธ์กับโนดอื่น ๆ เสมอ (no dependency) เช่น ในโครงสร้างความรวม (coordinated structure) ที่มีคำเชื่อม "และ" จะกำหนดให้หน่วยที่อยู่ข้างหน้าเป็นหน่วยหลัก

**3. ระบบภาษากลาง** ระบบภาษากลาง เป็นระบบที่พัฒนาขึ้นเพื่อเป็นรูปแทนความหมายของประโยคอินพุต โครงสร้างภาษากลางเป็นโครงข่ายที่ประกอบด้วยโนด มโนทัศน์ของหน่วยคำ และความสัมพันธ์ระหว่างมโนทัศน์ เช่น ความสัมพันธ์แบบ agent, object, goal เป็นต้น

**4. ระบบการจัดการของตัววิเคราะห์ภาษาทางคอมพิวเตอร์ (Computer System Organization)** โครงสร้างของตัววิเคราะห์ภาษาทางคอมพิวเตอร์ มีองค์ประกอบที่สำคัญ 2 ส่วน คือ ความรู้(knowledge) และตัวประมวลผลความรู้(software engine) ความรู้ที่ใช้ในการวิเคราะห์ประกอบด้วย พจนานุกรมของการวิเคราะห์ ตารางข้อมูล และฐานความรู้ทางคอมพิวเตอร์ ตัววิเคราะห์ภาษาที่คณะผู้เขียนกำลังดำเนินดารวิจัยอยู่อธิบายด้วยกฎซึ่งเขียนด้วยภาษาPIVOT




รูปที่ 3 องค์ประกอบของตัววิเคราะห์ภาษา

**4.1 โครงสร้างของฐานความรู้ทางคอมพิวเตอร์** ในระบบPIVOT ฐานความรู้ที่ใช้ในการวิเคราะห์จึงสร้างขึ้นโดยอาศัยไวยากรณ์โครงสร้างวลี ไวยากรณ์พึ่งพา และไวยากรณ์การก จะแสดงในรูปของกฎ รูปแบบของกฎเป็นproduction system กล่าวคือ กฎต่างๆจะแบ่งเป็น 2 ส่วนคือส่วนที่บอกเงื่อนไข และส่วนที่บอกการกระทำ ลำดับของการใช้กฎและเนื้อหาในส่วนที่บอกเงื่อนไข กับการกระทำแยกออกเป็น 2 แฟ้มคือ

ก. แฟ้มของการลำดับกฎ (Link File )

ข. แฟ้มของกฎ (Rule File)

**4.1.1 แฟ้มของการลำดับกฎ** จะอธิบายลำดับของการใช้กฎต่างๆ รูปแบบของการเขียน แบ่งออกเป็น 6 แบบดังนี้

แบบที่ 1 เงื่อนไขเดี่ยว - การกระทำเดี่ยว

กฎ : ถ้า Condition เป็นจริง แล้ว Action เกิด

รูปแบบ : { Condition – Action }

แบบที่ 2 เงื่อนไขเดี่ยว - การกระทำแบบ AND

กฎ : ถ้า Condition เป็นจริง แล้ว Action และ Action 2... Action เกิด

รูปแบบ : {Condition -- Action 1 | Action 2 | ... |Action n}

แบบที่ 3 เงื่อนไขหลายระดับ - การกระทำเดี่ยว

กฎ : ถ้า Condition 1 และ Condition 2 ... และ Condition n เป็นจริงตามลำดับ แล้ว action 1 เกิด

รูปแบบ : {Condition 1--

{Condition 2 --

...

{Condition n -- Action 1 }

แบบที่4 เงื่อนไขหลายระดับ - การกระทำแบบ AND

กฎ : ถ้า Condition 1 และ Condition 2 ... และ Condition n เป็นจริงตามลำดับแล้ว Action 1 และ Action 2 ... และ Action n เกิด

รูปแบบ : {Condition 1 --

{Condition 2 --

...

{Condition n -- Action 1 | Action 2 ...

| Action n }

แบบที่ 5 เงื่อนไขแบบลูกโซ่

กฎ : ถ้า Condition 1 เป็นจริง แล้ว Action 1 เกิด และ
ถ้า Condition 2 เป็นจริง แล้ว Action 2 เกิด และ

...

ถ้า Condition n เป็นจริง แล้ว Action n เกิด ตามลำดับ

รูปแบบ : {Condition 1 -- Action 1

{Condition 2 -- Action 2

...

{Condition n -- Action n}

}

}

แบบที่ 6 เงื่อนไขแบบกระโดด

กฎ : ถ้า Condition 1 เป็นจริง แล้วข้ามไปตรวจสอบกฎที่ชื่อ Label j

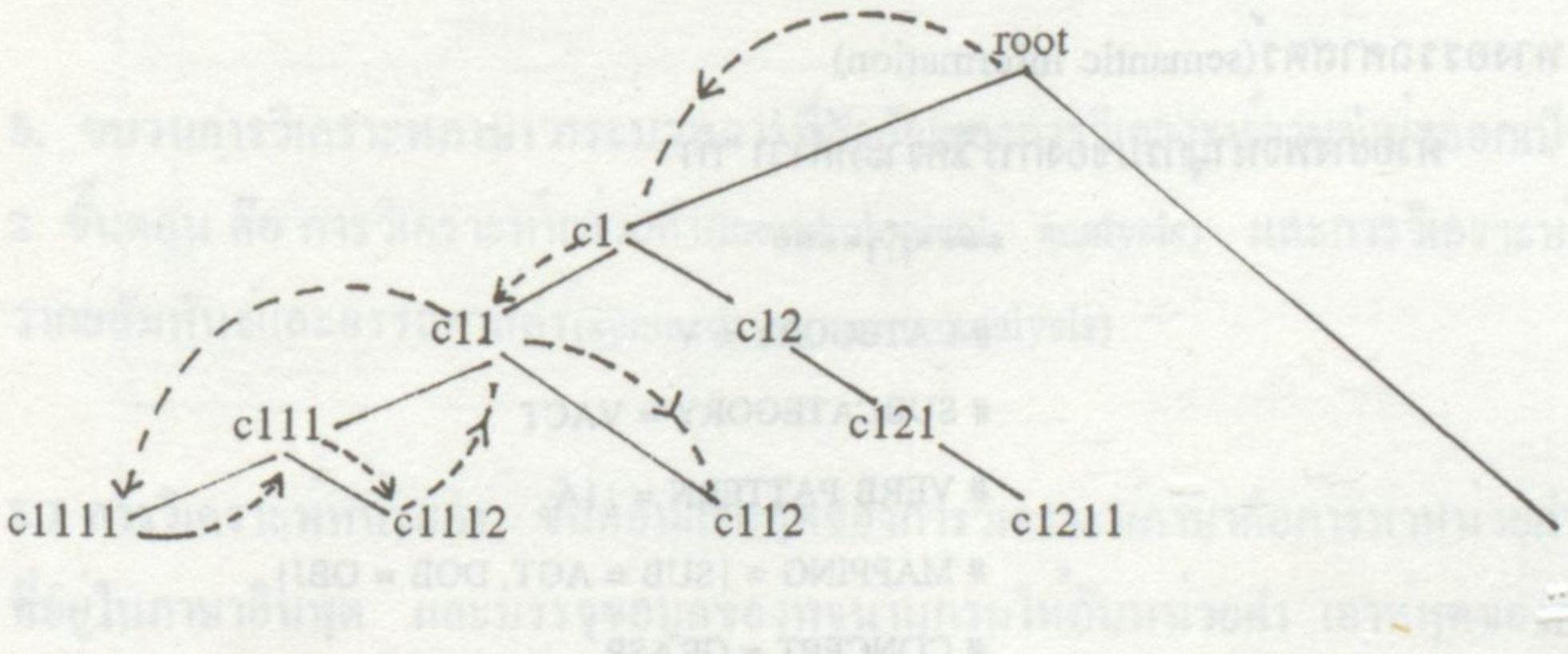
และเมื่อสิ้นสุดกฎแล้ว จะกลับไปที่กฎข้อแรก

รูปแบบ : {Condition 1-- (Label j)}

...

{<Label j> Condition j; -- Action j;}

การเรียงลำดับของกฎมีรูปแบบเป็นโครงสร้างต้นไม้ การเคลื่อนที่ (traverse) ของกฎบนต้นไม้นี้ เป็นไปตามหลักการของ depth - first search (ดังรูป 2)




รูปที่ 4 วิถีการเคลื่อนที่แบบ depth - first search

**4.1.2 แฟ้มของกฎ** รายละเอียดของเงื่อนไขและการกระทำจะเขียนไว้ในแฟ้มของกฎ ตัวอย่างต่อไปนี้แสดงรูปแบบของการเขียนรายละเอียดของกฎ

<Condition>

[

# operation c1; arg1, arg2, ..., argn > output c1

# operation c2; arg1, arg2, ..., argn > output c2

...

# operation cn: arg1, arg2, ..., argn > output c3

]

<Action>

[

# operation a1; arg1, arg2, ... , argn > output a1

# operation a2; arg2, arg2, ... , argn > output a2

...

# operation an; arg1, arg2, ... , argn > output an

]

**4.2 พจนานุกรมของการวิเคราะห์** เป็นฐานข้อมูลที่ใช้ในการวิเคราะห์ภาษาอินพุตซึ่งประกอบด้วยข้อมูลทางวากยสัมพันธ์(syntactic information) และข้อมูลทางอรรถศาสตร์(semantic information)

ตัวอย่างพจนานุกรมของการวิเคราะห์คำว่า "กำ"

****กำ****

# CATEGORY = V

# SUBCATEGORY = VACT

# VERB PATTERN = 11A

# MAPPING = {SUB = AGT, DOB = OBJ}

# CONCEPT = GRASP

# AKO = 2111

**********

# CATEGORY = CLAS

# SUBCATEGORY = CVBL

# CONCEPT = GRASP

# AKO = 2364

**********

**4.3 ตารางข้อมูล** นอกจากอาศัยข้อมูลจากพจนานุกรมของการวิเคราะห์แล้ว การวิเคราะห์ ยังต้องอาศัยข้อมูลเกี่ยวกับคุณสมบัติของกริยารูปแบบต่างๆ ซึ่งแสดงในรูปของตาราง ดังนี้

| verb pattern | category | voice Active | voice passive | Tense | Position | Prep spelling | Argument |
|---|---|---|---|---|---|---|---|
| 1 | V | 1 | 0 | 1 | 1 | - | SUBST |
| 5 | V | 1 | 1 | 1 | 1 | - | SUBST |
| | | | | | 1 | CMP3 | SUBST |
| 14 | V | 1 | 0 | 1 | 1 | - | SUBST |
| | | | | | 1 | CMP2 | THAT |

รูปที่ 4 ตารางข้อมูลรูปแบบบังคับเกิดของคำกริยา

**5. ขบวนการวิเคราะห์ภาษา** กระบวนการที่สำคัญของการวิเคราะห์ภาษาแบ่งออกเป็น 2 ขั้นตอน คือ การวิเคราะห์หน่วยคำ(morphological analysis) และการวิเคราะห์วากยสัมพันธ์และอรรถศาสตร์(syntactic-semantic analysis)

**5.1 การวิเคราะห์หน่วยคำ** ขั้นตอนแรกสุดของการวิเคราะห์ภาษาคือการหาหน่วยคำที่อยู่ในภาษาอินพุต และบรรจุข้อมูลของพจนานุกรมให้กับหน่วยคำ เอาท์พุตของการวิเคราะห์คือโครงสร้างการเรียงอันดับเชิงเส้นของหน่วยคำในประโยคอินพุต เช่น

ตัวอย่างประโยค

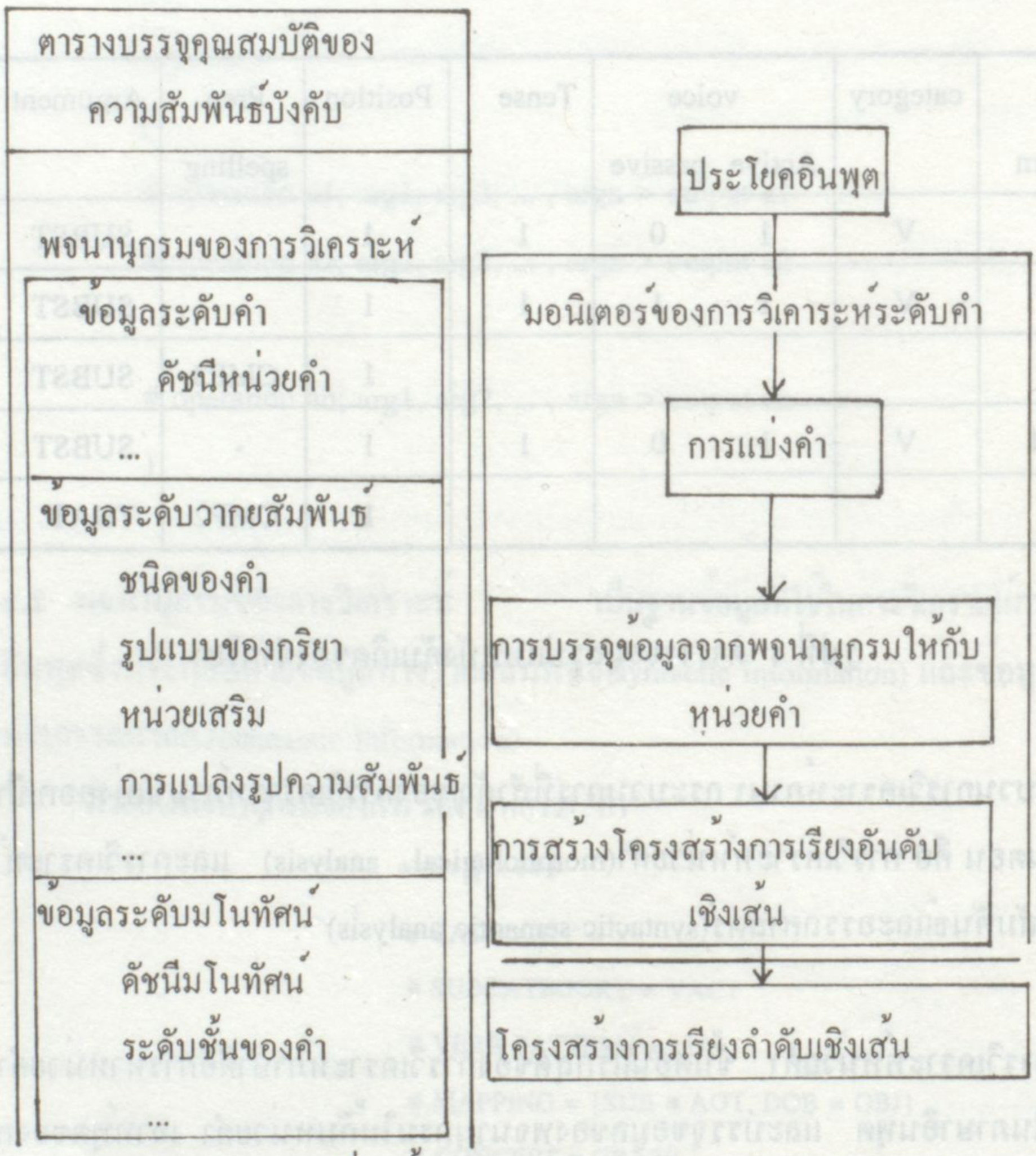
อินพุต ปัญญาประดิษฐ์เป็นวิชาใหม่

เอาท์พุต




รูปที่ 5 โครงสร้างการเรียงอันดับเชิงเส้นของหน่วยคำในประโยคอินพุต

การประมวลผลระดับคำแบ่งเป็น 3 ขั้นตอนดังที่แสดงให้เห็นในแผนภูมิต่อไปนี้




รูปที่ 6 ขั้นตอนการประมวลผลระดับคำ

**5.2 การวิเคราะห์วากยสัมพันธ์และอรรถศาสตร์** ขั้นตอนนี้ เป็นการวิเคราะห์ความสัมพันธ์ทางวากยสัมพันธ์และอรรถศาสตร์ระหว่างหน่วยคำต่างๆ โดยทั่วไปการหาความสัมพันธ์ระหว่างคำนามกับคำนามสามารถกำหนดความสัมพันธ์ ทางความหมายได้โดยตรง ส่วนความสัมพันธ์ระหว่างคำกริยากับคำนามต้องผ่านขบวนการวิเคราะห์ทางวากยสัมพันธ์ก่อนเสมอ ด้วยเหตุนี้ระบบการแปลภาษาหลายระบบในปัจจุบันจึงมิได้แยกขั้นตอนของการวิเคราะห์วากยสัมพันธ์และ อรรถศาสตร์ออกจากกัน

การวิเคราะห์ความสัมพันธ์ทางวากยสัมพันธ์และอรรถศาสตร์ จะช่วยลดความกำกวมทางไวยากรณ์ของรูปคำได้อีกด้วย ความรู้ที่ใช้ในการวิเคราะห์ประกอบด้วยพจนานุกรม ตารางแสดงองค์ประกอบของกริยา ตารางแสดงคุณสมบัติการบังคับเกิดของกริยา และฐานความรู้ทางวากยสัมพันธ์และอรรถศาสตร์ โดยอาศัยขั้นตอนดังนี้

ก. ตรวจสอบว่ามโนทัศน์ใดมีความสัมพันธ์กัน และมโนทัศน์ที่สัมพันธ์กัน มโนทัศน์ใดเป็นหลักและหน่วยใดเป็นหน่วยพึ่งพา การตรวจสอบนี้กระทำโดยอาศัยไวยากรณ์พึ่งพา และไวยากรณ์โครงสร้างวลี

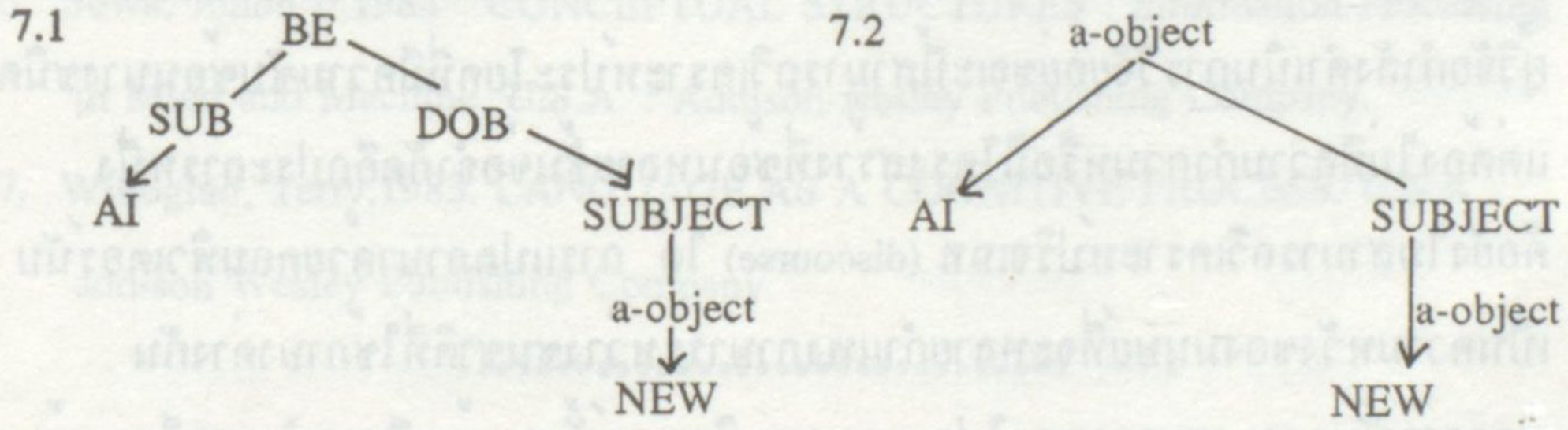
ข. เชื่อมมโนทัศน์ต่าง ๆ เข้าด้วยกัน แล้วกำหนดความสัมพันธ์

ค. รวมอาร์ก (arc) ของหน่วยหลัก และหน่วยพึ่งพาให้เป็นอาร์กเดียวกัน จะได้โครงสร้างพึ่งพาย่อย

4. ดำเนินการซ้ำตั้งแต่ขั้นตอน (1) ถึง (3)

เอาท์พุตของการวิเคราะห์ คือโครงสร้างต้นไม้ที่ประกอบด้วยโนดของต่าง ๆ และอาร์กของความสัมพันธ์ระหว่างคำ เช่น ประโยคข้างบน จะได้เอาท์พุตดังนี้

ปัญญาประดิษฐ์เป็นวิชาใหม่




**รูปที่ 7 ผลการวิเคราะห์วากยสัมพันธ์และอรรถศาสตร์**

จากรูปที่ 7.1 จะเห็นว่ามีการรวม วิชา(SUBJECT) กับ ใหม่(NEW)ให้เป็นนามวลีเสียก่อน โดย *ใหม่* ซึ่งมีคุณสมบัติทางอรรถศาสตร์เป็น attribute เป็นหน่วยพึ่งพาของ *วิชา* ดังนั้นจึงกำหนดการกสัมพันธ์เป็น a-object (จากนั้นจึงหาความสัมพันธ์ระหว่าง *เป็น* ซึ่งเป็นคำกริยากับคำนามที่ยังไม่เป็นหน่วยพึ่งพาของหน่วยใดๆ ซึ่งในที่นี้คือ*ปัญญาประดิษฐ์* เป็นประธาน(SUB) และ *วิชา* เป็นกรรมตรง(DOB)ของเป็น(รูปที่ 7.1) จากนั้นใช้สูตรของการแปลงความสัมพันธ์ทางวากยสัมพันธ์อรรถศาสตร์ของกริยา เป็นเปลี่ยนการกทางวากยสัมพันธ์ ให้เป็นการกทางอรรถศาสตร์ (รูปที่ 7.2)

**6. บทสรุป** การวิเคราะห์ภาษาไทยเป็นงานที่มีความาสำคัญเพิ่มขึ้นเป็นลำดับโดยเฉพาะอย่างยิ่ง ในยุคที่คอมพิวเตอร์เข้ามามีบทบาทในทุก ๆ ด้าน การวิเคราะห์ภาษาไทยจะเป็นประโยชน์ในการพัฒนาเทคโนโลยีของการสื่อสารระหว่างคนกับเครื่อง และการพัฒนาระบบการแปลภาษาด้วยคอมพิวเตอร์คณะผู้วิจัยได้พยายามสร้างฐานความรู้เพื่อให้คอมพิวเตอร์สามารถวิเคราะห์ความหมายของประโยคภาษาไทยให้อยู่ในรูปโครงข่ายทางความหมายที่เป็นสากล แต่อย่างไรก็ตามการขจัดรูปภาษาเฉพาะนั้นเป็นปัญหาใหญ่ ในการวิเคราะห์ด้วย เหตุที่ลักษณะไวยากรณ์ภาษาไทยเป็นภาษาคำโดดไม่มีวิภัติปัจจัย จึงทำให้เป็นภาษาที่มีความกำกวมสูงมาก ประกอบไปด้วยคำหลายหน้าที่ และคำหลายความหมายมากมาย นอกจากนั้นยังมีโครงสร้างที่มีกริยาเกิดเรียงกันได้หลายกริยา(serial verb) ที่สำคัญที่สุดไม่มีรูปบอกการสิ้นสุดประโยคหรืออนุประโยค ลักษณะดังกล่าวนี้จะต้องนำมาวิเคราะห์เพื่อหาข้อสรุปทางภาษาต่อไป ปัจจุบันความสามารถของตัววิเคราะห์ภาษาเพื่อการแปลภาษาด้วยคอมพิวเตอร์ที่คณะผู้วิจัยกำลังดำเนินการวิจัยอยู่ขณะนี้สามารถวิเคราะห์ประโยคที่มีความซับซ้อนบางชนิดแต่ต้องไม่มีความกำกวมหรือมีโครงสร้างที่ซ้อนหลายชั้นข้อจำกัดอีกประการหนึ่งคือยังไม่สามารถวิเคราะห์ปริจเฉท (discourse) ได้ การแปลภาษาด้วยคอมพิวเตอร์นับเป็นความหวังของมนุษย์ที่จะทลายกำแพงภาษาระหว่างชนชาติที่ใช้ภาษาต่างกัน นอกจากนั้นความสามารถของโปรแกรมคอมพิวเตอร์นั้นจะต้องมีการทำงานเร็วถูกต้องและแม่นยำเพื่อจะได้เป็นเครื่องแปลภาษาที่มีประสิทธิภาพ

**เชิงอรรถ**

*คณะนักวิจัยที่ทำงานโครงการวิเคราะห์ภาษาไทยเพื่อการแปลภาษาด้วยคอมพิวเตอร์ประกอบไปด้วย

1. ผศ. วรนุช เกิดสินธุ์ชัย อาจารย์ภาควิชาคณิตศาสตร์ คณะครุศาสตร์วิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้า ธนบุรี
2. สุดา รังกุพันธุ์ อาจารย์ภาควิชาภาษาอังกฤษ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
3. ทัศนีย์ เจริญพร นักวิจัยห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ ศูนย์เทคโนโลยีอิเลคโทรนิกส์และคอมพิวเตอร์แห่งชาติ

**REFERENCES**

1. Chafe,Wallace L.1970. MEANING AND THE STRUCTURE OF LANGUAGE. U.S.A : The University of Chicago Press.
2. Chomsky,Noam. 1965. ASPECT OF THE THEORY OF SYNTAX . U.S.A. : The MIT Press.
3. Dowty, David R., Lauri Karttunen, and Arnolod M. Zwicky. 1985. : NATURAL LANGUAGE PARSING.U.S.A. : Cambridge University Press.
4. Fillmore, C.J,1968. The case for case, In Bach and Harms (1968: 1-88)
5. Melcuk,lgorA.1988. DEPENDENCYU SYNTAX: THEORY AND PRACTICE. Albany : State of University of new York Press.
6. Sowa, Johbn F,1984 CONCEPTUAL STRUCTURES : Information Processing in Mind and Machine. U.S.A. : Addison-wesley Publishing Company.
7. Winograd, Terry,1983. LANGUAGE AS A COGNITIVE PROCESS. U.S.A : addison Wesley Publishing Company.

**ภาคผนวก**

การแบ่งประเภทของคำ แบ่งเป็น 13 ประเภทดังนี้คือ

NOUN

| | | |
|---|---|---|
| NPRP | proper name | กรุงเทพฯ มานะ ไทย |
| NCNM | cardinal number | หนึ่ง สอง สาม |
| NONM | ordinal number | ที่หนึ่ง ที่เท่าไร |
| NALN | adverb like noun | วันอาทิตย์ วันนี้ |
| NATN | attributive noun | น้ำเงิน แดง ร้อน |
| NLBL | label | 1,2,3,..... |
| NCMN | common noun | หนังสือ อาคาร หอ การทำงาน |
| NTTL | title | ร้อยเอก ดร. |
| NNOM | nominalized noun | การทำงาน ความรัก |
| NCTV | collective noun | หมู่ กอง โขลง |

PRONOUN

| | | |
|---|---|---|
| PPRS | personal pronoun | กระผม ข้าพเจ้า เขา พวกเรา |
| PDMN | demonstrative pronoun | นี่ นั่น นี้ |
| PDEF | definite pronoun | เหล่านี้ ทั้งหมด ทั้งคู่ ทั้งสอง |
| PIND | indefinite pronoun | บ้าง ทั้งหลาย ทั้งหมด ทั้งมวล |
| PNTR | interrogative pronoun | ใคร อะไร ที่ไหน อย่างไร เมื่อไร ทำไม |
| PREL | relative pronoun | ที่ ซึ่ง อัน ผู้ ที่ว่า |

VERB

| | | |
|---|---|---|
| VACT | active verb | บรรยาย เสนอ ดู ฟัง |
| VSTA | stative verb | ราคา ได้ยิน เห็น เสียใจ |
| VATT | attributive verb | สวย ใหญ่ ผอม หนัก |
| VEQU | equative verb | เป็น อยู่ คือ ได้แก่ ชื่อ เท่ากับ |

| | | |
|---|---|---|
| | | หมายถึง |
| VCAU | causative verb | ทำ ทำให้ ให้ |
| VMTN | motive verb | กระโจน จ้ำ บิน มา วิ่ง กระโดด เดิน ไป ย่อง คลาน ทะยาน เผ่น เลื้อย (นั่ง นอน ยืน) |
| AUXILIARY | | |
| XVBM | before verb and before "ไม่" in negation | จะ กำลัง คล้ายจะ อาจ เกือบ |
| XVAM | before verb and after "ไม่" in negation | ค่อย น่า ได้ มัว |
| XVMM | before verb and before or after "ไม่" in negation | ควร เคย ต้อง อยาก ชอบ |
| XVAE | after verb;at the end of verb phrase | อยู่ แล้ว อยู่แล้ว ไป มา ขึ้น ลง เข้า ออก เสีย ไว้ เอา ให้ ดู |
| DETERMINER | | |
| DDAN | after noun without clas, in between | ทั้งหลาย ทั้งหมด ส่วนใหญ่ |
| DDAC | after noun allow clas, in between | นี่ นี้ นั่น นั้น ไหน ใด อื่น |
| DDBQ | between noun and clas, or before quantitative expression | ทั้ง ตั้ง สัก ทุก เพียง เกือบ จวน ราว แค่ กว่า รวม รวม ถึง ใกล้ นับ เกิน เลย เฉียด เหยียบ |
| DDAQ | after quantitative expression | เท่านั้น พอดี ถ้วน ครึ่ง เศษ กว่า กว่า ๆ |

| | | |
|---|---|---|
| DCNM | cardinal number experssion | หนึ่ง เก้า สาม เดียว |
| DONM | ordinal number expression | ที่หนึ่ง ที่สอง แรก สุดท้าย หน้า กลาง หลัง |
| ADVERB | | |
| ADVN | adverb with normal form | จำเจ เก่ง เร็ว |
| ADVI | adverb with iterative form | เร็วๆ เสมอๆ |
| ADVS | sentential adverb | โดยปกติ ทั่วๆ ไป ตามธรรดา |
| ADVT | interrogative adverb | ทำไม เท่าไร อย่างไร |
| CLASSIFIER | | |
| CNIT | unit classifier | คน ตัว ต้น อัน ข้อ ประการ |
| CLTV | collective classifier | กลุ่ม ฝูง คณะ เครือ โขลง |
| CMTR | metric classifier | เมตร ลิตร ซม. วัน ชั่วโมง แก้ว ช้อน ถุง ถัง |
| CFQC | frequency classifier | ครั้ง ที่ หน เที่ยว รอบ |
| CVBL | verbal classifier | ขด จีบ มัด มวน ม้วน |
| PREPOSITION | | |
| RVBP | verbal preposition | จาก ข้าม ละ |
| RPRE | preposition except "RVBP" | ก่อน ใต้ บน .... |
| CONJUNCTION | | |
| CCRG | coordination conjunction | และ หรือ แต่ แต่ว่า |
| CCMP | comparative conjunction | กว่า ที่สุด มากกว่า |
| CSBR | subordinating conjunction | ได้แก่ ว่า ที่ ตามที่ว่า อย่างที่ ซึ่ง อัน ตัว ผู้ ดังที่ เนื่องจาก ระหว่าง... |
| ENDING | | |
| EAFF | ending for affirmative | ครับ ค่ะ นะ จ้ะ น่ะ ล่ะ |

| | | |
|---|---|---|
| | sentence | ฮ่ะ ฮะ |
| EITT | ending for interrogative sentence | หรือ เหรอ ไหม ฮะ คะ จ๊ะ |
| EIMP | ending for imperative sentence | ซิ ซิ่ ซี่ เถอะ น่า หน่อย |
| INTERJECTION | | |
| INT | interjection | โอ้ย โอ้ เออ อ้อ |
| PREFIX | | |
| PIXP | prefix | อย่าง โดย |
| NEGATION | | |
| NEG | negation | ไม่ได้ ไม่ มิได้ มิ |

# การตัดคำไทยในระบบแปลภาษา

# Word Segmentation for Thai in Machine Translation System

วิรัช ศรเลิศล้ำวาณิช

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้

ศูนย์เทคโนโลยีอิเลคทรอนิกส์และคอมพิวเตอร์แห่งชาติ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

**1. บทนำ** การตัดคำ หรือการแบ่งข้อความที่ต่อเนื่องกันออกเป็นหน่วยคำๆ (Morpheme) หรือลักษณะของการรู้จำ (Recognition) คำหนึ่งๆในข้อความที่ต่อเนื่องกันนั้นเริ่มจะมีความหมายมากยิ่งขึ้นเป็นลำดับเมื่อมีการนำเอาคอมพิวเตอร์เข้ามาช่วยในการประมวลผลข้อมูลทางภาษามากยิ่งขึ้น ความยากง่ายหรือวิธีการที่จะนำมาใช้ในการตัดคำนั้นต้องขึ้นอยู่กับลักษณะเฉพาะของภาษานั้นๆเป็นอย่างมาก ผลงานวิจัยที่นำมาเสนอนี้ไม่ได้เป็นคำตอบสุดท้ายหรือคำตอบที่ดีที่สุด แต่เป็นวิธีที่ใช้ได้ผลดีและก็กำลังใช้อยู่ในการพัฒนาระบบเครื่องแปลภาษา ผลของการใช้งานจริงได้เป็นตัวกระตุ้นการวิจัยและพัฒนาระบบการตัดคำนี้จนได้เป็น Version 2 ในปัจจุบันนี้

**2. จุดประสงค์ของการตัดคำ** การตัดคำหรือที่เรียกกันว่า Word Segmentation นั้นคือการแบ่ง String ของตัวอักษรเพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) เนื่องจากว่าโดยปกติทั่วไปแล้ว ภาษาไทยมีการเขียนในลักษณะที่ติดต่อกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใดๆ ยกเว้นแต่มีการเว้นวรรคเป็นระยะๆ เพื่อให้ผู้อ่านได้หยุดพัก และทำความเข้าใจความหมายเป็นตอนๆไปเท่านั้น แม้ว่าการเว้นวรรคในการเขียนบทความไม่ได้มีกฎเกณฑ์ที่ชัดเจนก็ตาม แต่ถ้ามีการใช้การเว้นวรรคด้วยความระมัดระวังแล้วก็จะสามารถช่วยลดความคลุมเครือของคำหรือประโยคได้

ไม่ว่าจะด้วยวิธีการใดก็ตามถ้าหากสามารถรู้เขตแบ่งของแต่ละคำได้แล้ว การจัดการกับข้อความนั้นๆก็จะเป็นไปได้อย่างสะดวกและถูกต้อง ในระบบคอมพิวเตอร์จึงจำเป็นอย่างยิ่งที่จะต้องคำนวณหาขอบเขตของแต่ละคำให้ได้เพื่อที่จะสามารถลดภาระ

ของผู้ใช้หรือเพื่อที่จะให้กระบวนการที่อยู่ในระดับที่ลึกกว่าสามารถทำงานต่อไปได้ ดังเช่น ฟังก์ชั่นการขอบขวา (Word wrapping) ใน Word processor การตรวจคำผิด การค้นหาคำใน text หรือเป็นตัวช่วยในการกำหนดคำเพื่อทำการวิเคราะห์ต่อไปในระบบของเครื่องแปลภาษา

**3. อัลกอริธึมสำหรับการตัดคำโดยใช้พจนานุกรม** การตัดคำคือการหาขอบเขตของหน่วยคำในข้อความที่ต่อเนื่อง ฉะนั้นถ้าหากเก็บทุกคำที่มีอยู่ในภาษาลงในพจนานุกรมทั้งหมด จากนั้นก็ค้นหาและเปรียบเทียบหาคำศัพท์นั้นๆว่ามีอยู่ในพจนานุกรมหรือไม่ เพียงเท่านั้นก็จะสามารถหาขอบเขตของคำแต่ละคำได้ แต่ในความเป็นจริงแล้วไม่สามารถจะกระทำได้ เนื่องจากว่าเป็นไปไม่ได้ที่จะบรรจุคำทุกคำลงในพจนานุกรมได้ทั้งหมด โดยเฉพาะในส่วนที่เป็นวิสามานยนาม (คำนามที่เป็นชื่อเฉพาะ---พระยาอุปกิตศิลปสาร, 1990) ตัวเลข หรือคำที่เกิดขึ้นมาจากการใช้ใหม่ๆ หรือมีการบัญญัติขึ้นมาใหม่ ซึ่งไม่สามารถจะคาดการล่วงหน้าได้ ฉะนั้นการตัดคำถึงแม้ว่าจะอาศัยการเปรียบเทียบคำจากพจนานุกรมก็ตาม ก็จำเป็นที่จะต้องยอมให้มีคำที่ไม่ได้บรรจุไว้ในพจนานุกรมเกิดขึ้นได้เช่นกัน

คำที่จะบรรจุอยู่ในพจนานุกรมไม่จำเป็นที่จะต้องเป็นหน่วยคำที่ย่อยที่สุดที่คงความหมายไว้เสมอไป อาจเป็นคำประสม (แม่น้ำ, ดูแล, ช่างทอง, เป็นต้น) หรือวลี (แสงอาทิตย์---รัศมี+ดวงตะวัน, การแปลภาษาด้วยเครื่องคอมพิวเตอร์, หนีเสือปะจรเข้, เป็นต้น) ก็ได้ ทั้งนี้ก็ขึ้นอยู่กับว่าตำแหน่งของคำในวลีและโครงสร้างทางวากยสัมพันธ์นั้นๆคงที่แน่นอนหรือไม่ และจะสามารถกำหนดความหมายที่ชัดเจนให้กับวลีนั้นๆได้หรือไม่

**3.1 กฎทางอักขรวิธี** แม้ว่าการเขียนข้อความในภาษาไทยจะไม่มีการเว้นวรรคระหว่างคำ ไม่มีการใช้เครื่องหมายวรรคตอนที่ชัดเจน ไม่มีตัวชี้บ่งหน้าที่ทางไวยกรณ์ (Syntactic Marker---Postposition) ไม่มีการแปรรูป (Inflection) ไม่มีการใช้ตัวอักษรใหญ่-เล็ก เป็นต้น แต่ภาษาไทยก็มีกฎทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรคดังได้กล่าวไว้ในหัวข้อที่ 2. และการขึ้นย่อหน้า ซึ่งทั้ง 3 ลักษณะนี้

จะเป็นตัวช่วยในการชี้บ่งขอบเขตของการพิจารณาในการเปรียบเทียบกับคำในพจนานุกรม

การขึ้นย่อหน้าเป็นตัวชี้บ่งได้ดีถึงการสิ้นสุดของความ

การเว้นวรรคเป็นตัวชี้บ่งถึงความเป็นไปได้ของการสิ้นสุดของคำหรือประโยค

กฎทางอักขรวิธีเป็นตัวชี้บ่งถึงความเป็นไปได้ของการพิจารณาในการที่จะแยกสายอักขระออกเพื่อการพิจารณาคำหรือไม่ ฉะนั้นจึงมีการแบ่งอักขระออกใหม่ดังนี้

อักขระกลุ่มที่ 1 กลุ่ม Non-spacing Character คือกลุ่มของรูปสระ วรรณยุกต์ และเครื่องหมายพิเศษประกอบการเขียน ที่เมื่อประสมเข้ากับพยัญชนะใดๆแล้วไม่ทำให้มีการเคลื่อนขวาของตำแหน่งที่จะเขียนต่อไป อักขระในกลุ่มนี้จะไม่สามารถอยู่เดียวๆ ได้ มีดังนี้คือ ั ิ ี ึ ื ุ ู ็ ่ ้ ๊ ๋ ์ ๎

อักขระกลุ่มที่ 2 กลุ่มอักขระที่จำเป็นที่จะต้องมีรูปพยัญชนะตามเสมอ มีดังนี้คือ เ_, แ_, โ_, ใ_, ไ_

อักขระกลุ่มที่ 3 กลุ่มอักขระที่จำเป็นที่จะต้องมีรูปพยัญชนะอยู่หน้าเสมอ มีดังนี้คือ _ะ, _า, _ำ, _ๅ สำหรับ _ๅ จะมีพยัญชนะนำหน้าได้อยู่เพียง 2 ตัวคือ ฤ กับ ฦ

อักขระกลุ่มที่ 4 กลุ่มของอักขรที่เป็นตัวการันต์ที่มีไม้ทัณฑฆาตบังคับข้างบน เช่น ย์ ใน ทิพย์, น์ ใน กาญจน์ เป็นต้น เนื่องจากว่าตัวการันต์เป็นพยัญชนะสุดท้ายที่ไม่อ่านออกเสียง ฉะนั้นจะไม่มีการพิจารณาให้เป็นอักษรแรกของคำ

อักขระกลุ่มที่ 5 กลุ่มของอักขรที่เหลือทั้งหมด

เนื่องจากว่าข้อมูลสายอักขระถูกเก็บไว้ในระบบ X-TIS (รหัส 2 ไบท์สำหรับอักษรไทย) ดังนั้นอักขระในกลุ่มที่ 1 จึงไม่ต้องคำนึงถึง เพราะกลายเป็นส่วนหนึ่งของรหัสผสมไปแล้ว

อักขระในกลุ่มที่ 2, 3 และ 4 จะเป็นตัวจำกัดที่ดีที่จะไม่ให้มีการพิจารณาการตัดคำในระหว่างตัวอักษรนั้นๆ

**3.2 Longest Matching** วิธีการนี้ถือได้ว่าเป็นวิธีทาง Heuristic อันหนึ่ง โดยจะทำการตรวจสอบคำในพจนานุกรมแล้วแยกเป็นคำๆจากซ้ายไปขวา การตรวจสอบจะเริ่มตั้งแต่หน่วยของข้อความที่สั้นที่สุดที่จะสามารถคัดมาได้ เนื่องจากว่าภาษาไทยไม่มีเครื่องหมายบอกจุดสิ้นสุดของประโยค ดังนั้นจากเกณฑ์ทางอักขระดังในข้อที่ 3.1 จะเป็นตัว

บอกถึงขอบเขตของหน่วยข้อความได้ เมื่อตรวจสอบกับคำในพจนานุกรมแล้วถ้าไม่พบก็จะทำการลดความยาวของข้อความลงทีละตัวไปตามเกณฑ์ทางอักขรวิธี เช่น ในข้อความ "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ" เมื่อไม่ปรากฏในพจนานุกรมแล้วข้อความนี้ก็จะถูกลดลงเหลือ "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคั" แล้วก็เป็น "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำ" จนในที่สุดก็จะได้คำว่า "ความก้าวหน้า" เป็นคำแรก ผลของการตัดคำทั้งหมดจะเป็นดังในรูปที่ 1.

| ส่วนของคำที่ยาวที่สุด | ส่วนที่เหลือ |
|---|---|
| ความก้าวหน้า | ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ |
| ทาง | ด้านวิทยาศาสตร์มีบทบาทสำคัญ |
| ด้าน | วิทยาศาสตร์มีบทบาทสำคัญ |
| วิทยาศาสตร์ | มีบทบาทสำคัญ |
| มี | บทบาทสำคัญ |
| บทบาท | สำคัญ |
| สำคัญ | |

รูปที่ 1. แสดงผลของการตัดคำโดยวิธี Longest Matching ภายใต้เงื่อนไขของอักขรวิธี

วิธีการนี้จะให้ความถูกต้องในการตัดคำได้ประมาณ 80% ข้อบกพร่องที่เห็นได้ชัดเจนคือการเลือกขอบเขตของคำที่ยาวเกินไปตั้งแต่แรก จึงทำให้คำที่ตามมาผิดเพี้ยนไป เช่น ข้อความ "กีฬาเป็นการออกกำลังกายอย่างหนึ่ง" จะถูกแบ่งออกเป็น "กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง." เสมอ เพราะคำที่ยาวที่สุดที่จะพบได้ในพจนานุกรมสำหรับคำที่สองหลังจากที่ได้คำว่า "กีฬา" แล้ว จะเป็นคำว่า "เป็นการ" เสมอ ฉะนั้นจึงไม่สามารถจะแบ่งให้ได้คำว่า "ออกกำลังกาย" ที่มีความหมายที่ถูกต้องได้ในข้อความ

**3.3 วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรม (Unknown Word) น้อยที่สุด** นี่เป็นวิธีการทาง Heuristic อีกวิธีหนึ่ง และจะช่วยได้มากถ้าข้อความนั้นประกอบด้วยคำจำนวนมากหรือมีความยาวของข้อความมากพอสมควร วิธีนี้จะทำบน

อัลกอริธึมของ Longest Matching ในหัวข้อที่ 3.2 อีกทีหนึ่ง โดยจะทำการหาความเป็นไปได้ทั้งหมดในการตัดคำในข้อความนั้นๆ

Backtracking จะเริ่มกระทำหลังจากที่ได้คำตอบจากวิธี Longest Matching แล้ว โดยจะกระทำไปทีละคำจากซ้ายไปขวา เช่น "กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง." ซึ่งเป็นผลที่ได้จากการตัดแบบ Longest Matching จะถูก Backtrack ที่คำว่า "กีฬา" แต่เนื่องจากว่าจะไม่เกิดประโยชน์อันใดในการที่จะยอมให้เกิดคำที่ไม่มีในพจนานุกรมอีก ฉะนั้นการ Backtracking จึงสิ้นสุดลงตรงคำว่า "กีฬา" ต่อไปก็ทำการ Backtracking ที่คำว่า "เป็นการ" เพื่อตรวจดูความเป็นไปได้ทั้งหมดในการแบ่งคำ ผลที่ได้คือจะสามารถแบ่งได้เป็น "เป็น/การ...." และการ Backtracking ก็สิ้นสุดลงตรงคำว่า "เป็น"

เมื่อทำการ Backtracking สิ้นสุดลงกับทุกคำแล้ว ก็จะทำการคำนวณหา Cost ให้กับแต่ละ Path ที่เป็นไปได้ โดยบังคับให้มีการเกิด Unknown Word น้อยที่สุด แล้วเรียงผลของการตัดคำที่ได้ใหม่โดยให้คำตอบที่น่าจะเป็นไปได้มากที่สุดหรือถูกต้องมากที่สุดมาในอันดับแรก ผลของการตัดคำและ Cost ที่คำนวณได้แสดงไว้ในรูปที่ 2.

| ผลจากการ Backtracking | Cost |
|---|---|
| กีฬา/เป็น/การออกกำลังกาย/อย่างหนึ่ง. | 4 |
| กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง. | 5 |
| กีฬา/เป็นการ/ออก/กำลังกาย/อย่างหนึ่ง. | 5 |
| กีฬา/เป็นการ/ออกกำลัง/กาย/อย่าง/หนึ่ง. | 6 |
| กีฬา/เป็นการ/ออกกำลัง/กา/ยอ/ย่าง/หนึ่ง. | 7 |
| กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่างหนึ่ง. | 11 |
| กีฬา/เป็นการ/ออกกำลัง/กาย/อย่า/[ง]/หนึ่ง. | 12 |
| กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่าง/หนึ่ง. | 12 |
| กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่า/[ง]/หนึ่ง. | 18 |

รูปที่ 2. แสดงผลของการตัดคำโดยเรียงตาม Cost ที่คำนวณได้

**4. บทสรุป** การตัดคำเป็นกระบวนการเริ่มแรกที่จะนำข้อมูลเข้าสู่ระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์ และต้องการการประมวลผลที่มีความเร็วสูง วิธีดังกล่าวในงาน

วิจัยนี้แม้ว่าจะไม่สมบรูณ์ในตัวมันเอง กล่าวคือ ไม่สามารถที่จะให้ผลของการตัดคำที่ถูกต้องได้ 100% ก็ตาม เนื่องจากว่าในกระบวนการดังกล่าวไม่ได้ทำการวิเคราะห์ให้ลึกไปจนถึงโครงสร้างททางไวยากรณ์หรือความสัมพันธ์ทางความหมายเลย แต่ในเชิงปฏิบัติ เมื่อใช้หลักการทั้ง 3 ประการประกอบเข้าด้วยกันแล้ว ผลที่ได้จะให้ความถูกต้องได้สูงเกิน 90% เลยทีเดียว และในขณะเดียวกันก็ให้ผลของการตัดคำได้เร็วพอสำหรับการใช้งานในขั้นต่อๆไปได้ สำหรับระบบที่ต้องการความถูกต้องในการหาขอบเขตของคำสูงมากๆ ดังเช่น ในระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์ ระบบการค้นหาคำศัพท์ ระบบKWIC เป็นต้น ก็สามารถที่จะใช้ผลที่ได้ลองลงมาสำหรับการประมวลต่อไปได้

## เอกสารอ้างอิง


1. Tanaka Hozumi, 1989, Fundamental of Natural Language Analysis, Sangyotosho, in Japanese.
2. พระยาอุปกิตศิลปสาร, 1990, หลักภาษาไทย, ไทยวัฒนาพานิช

# การสังเคราะห์ภาษาไทยจากภาษากลางในระบบการแปลหลายหลายภาษา

ดร. รัตติกร วรากูลศิริพันธุ์
สมศักดิ์ จันวัน
อรัญญา ปรีชาไว
วิเชียร อัศวธีระ
ภาควิชาอิเล็กทรอนิกส์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

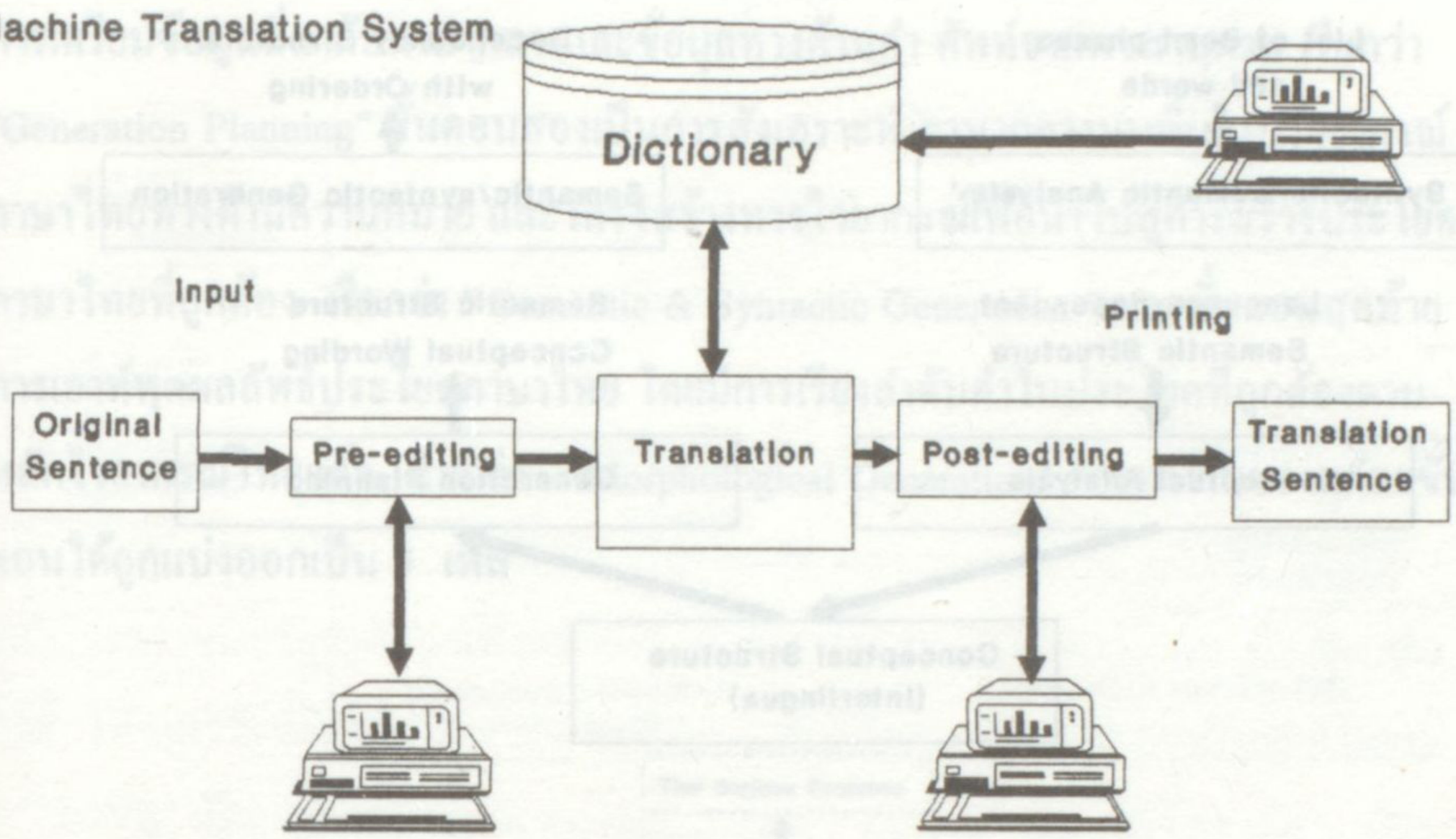
**บทคัดย่อ**

การแปลภาษาโดยใช้ภาษากลางเป็นตัวเชื่อมระหว่างภาษาต้นฉบับกับภาษาเป้าหมายนั้นเป็นกระบวนการแปลภาษาแบบถ่ายทอดทางอ้อมแบบหนึ่ง โดยแบ่งเป็น 2 ขั้นตอนใหญ่ๆ ได้แก่ขั้นตอนการวิเคราะห์ภาษาต้นฉบับไปสู่ภาษากลาง และขั้น ตอนการสร้างภาษาเป้าหมายจากภาษากลาง วิธีการแปลภาษาแบบนี้จะช่วยลดความยุ่งยากและลดเวลาในการวิเคราะห์ภาษาต้นฉบับและการสร้างภาษาเป้าหมาย เมื่อเปรียบเทียบกับกระบวนการแปลภาษาแบบถ่ายทอดโดยตรง ทั้งนี้เพราะเราไม่จำเป็นต้องคำนึงถึงคู่ภาษาที่จะแปล เนื่องจากการวิเคราะห์ภาษาต้นฉบับใดๆ ก็ตามให้เป็นภาษากลางแล้วก็จะสามารถแปลเป็นภาษาเป้าหมายอะไรก็ได้ โดยที่ไม่ต้องวิเคราะห์ใหม่และในทำนองเดียวกันทางด้านการสร้างภาษาเป้าหมายใดๆ นั้น ก็เป็นการสังเคราะห์จากภาษากลางซึ่งเป็นผลของการวิเคราะห์ภาษาต้นฉบับใดๆเพียงอย่างเดียว เมื่อมีการสังเคราะห์แล้วก็ไม่ต้องทำซ้ำอีกต่อไปนอกจากนั้นการแปล ภาษาโดยใช้ภาษากลางนี้จะใช้พจนานุกรมภาษากลางเป็นฐานข้อมูลในการวิเคราะห์ และสังเคราะห์เพียงชุดเดียวต่อหนึ่งภาษา ซึ่งต่างจากการแปลภาษาแบบถ่ายทอดโดยตรงที่จะต้องใช้พจนานุกรมถึง n ชุด ในการแปลจากภาษาต้นฉบับ 1 ภาษาไปสู่ภาษาเป้าหมาย n ภาษา

ในการสังเคราะห์ประโยคภาษาไทยจากภาษากลางของงานวิจัยนี้ ได้ศึกษาถึงโครงสร้างของภาษากลาง ซึ่งจะต้องถูกกำหนดขึ้นให้เป็นตัวแทนทางด้านความหมาย

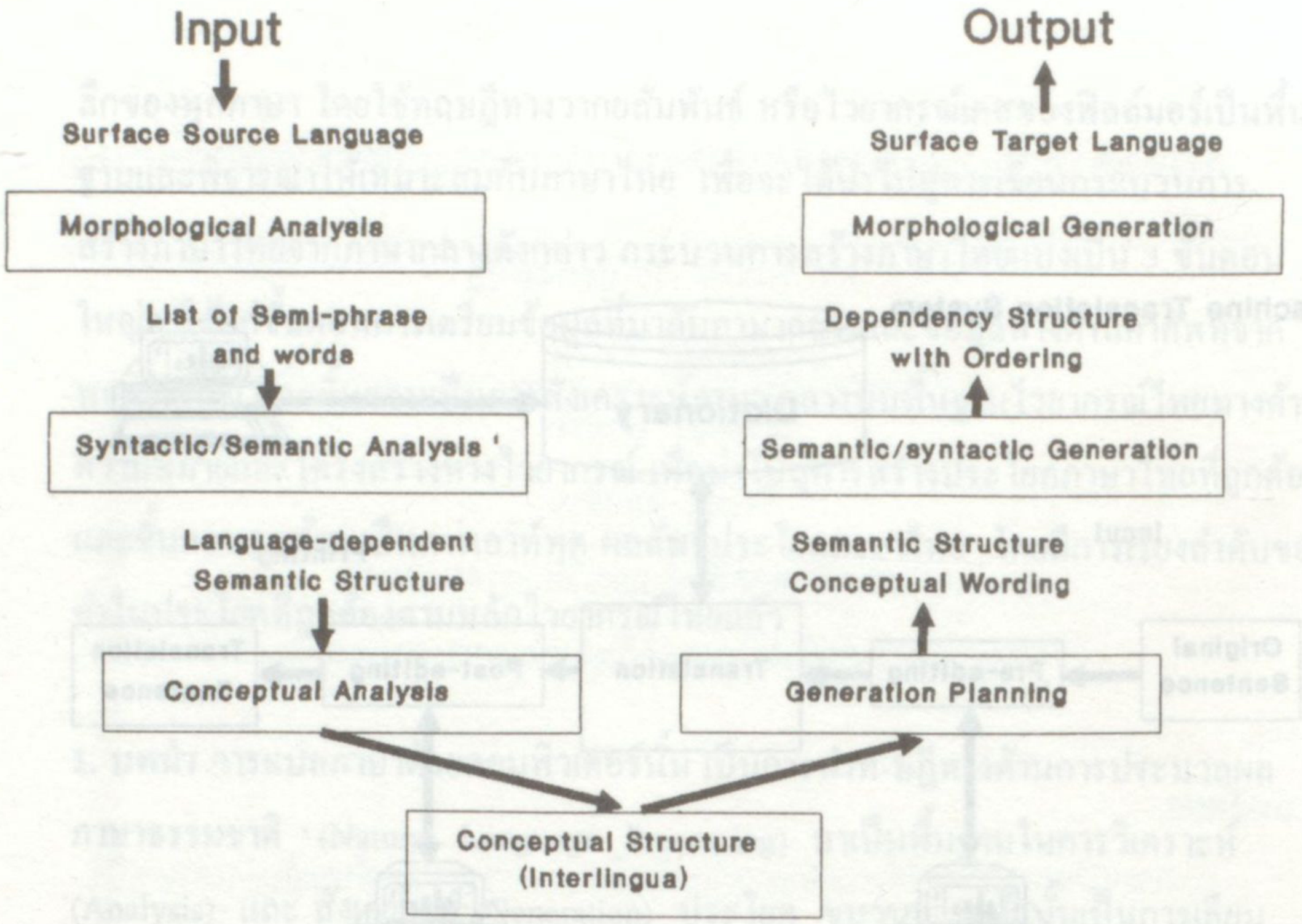
ลึกของทุกภาษา โดยใช้ทฤษฎีทางวากยสัมพันธ์ หรือไวยากรณ์เคสของฟิลล์มอร์เป็นพื้นฐานและพิจารณาให้เหมาะสมกับภาษาไทย เพื่อจะได้นำไปสู่การเขียนกระบวนการสร้างภาษาไทยจากภาษากลางดังกล่าว กระบวนการสร้างภาษาไทยแบ่งเป็น 3 ขั้นตอนใหญ่ๆ ได้แก่ขั้นตอนการเตรียมข้อมูลที่มากับภาษากลางและข้อมูลทางด้านคำศัพท์จากพจนานุกรม และขั้นตอนเป็นการสังเคราะห์ภาษา กลางบนพื้นฐานไวยากรณ์ไทยทางด้านความหมายและโครงสร้างทางไวยากรณ์ เพื่อนำไปสู่การสร้างประโยคภาษาไทยที่ถูกต้อง และขั้นตอนสุดท้าย เป็นการเอาท์พุต ผลลัพธ์ประโยคภาษาไทย โดยมีการเรียงลำดับของคำในประโยคที่ถูกต้องตามหลักไวยากรณ์ไทยแล้ว

**1. บทนำ** การแปลภาษาด้วยคอมพิวเตอร์นั้น เป็นการนำท-ยฎีทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาเป็นพื้นฐานในการวิเคราะห์ (Analysis) และ สังเคราะห์ (Generation) ประโยค ขบวนการแปลนั้นเป็นการเลียนแบบวิธีการแปลโดยธรรมชาติของบุคคล นั่นคือ จะต้องมีการอ่านประโยคของภาษาต้นฉบับ (Source Language) ที่จะแปล และทำความเข้าใจในความหมาย และโครงสร้างของประโยค ต่อจากนั้นก็ค้นหาคำศัพท์ของภาษาเป้าหมาย (Target Language) ที่มีความหมายตรงกับคำในภาษาต้นฉบับ แล้วเรียงคำให้ถูกต้องตามหลักไวยากรณ์ของภาษาเป้าหมายนั้น ซึ่งอาจจะต้องมีการตกแต่งประโยคให้ถูกต้องยิ่งขึ้น ดังนั้นปัญหาของการแปลภาษาด้วยเครื่องคอมพิวเตอร์จึงอยู่ที่ว่าจะสามารถทำให้คอมพิวเตอร์เข้าใจประโยคหรือข้อความที่จะแปลได้เหมือนมนุษย์ (ผู้แปล) อย่างไร ปัญหานี้สามารถแก้ไขได้ด้วยการสร้างฐานความรู้ (Knowledge Base) ในการแปลให้ระบบคอมพิวเตอร์ฐาน ความรู้ที่ว่านี้ได้แก่ ระบบการวิเคราะห์ ประโยค (Text Analysis System) และ ระบบการสังเคราะห์หรือสร้างประโยค (Text Generation System) ซึ่ง บรรจุไว้ด้วยกฎไวยากรณ์ของภาษา ตลอดจนพจนานุกรม (Electronic Dictionary) ของคู่ภาษาที่จะแปล ดังแสดงไว้ในรูปที่ 1




รูปที่ 1 โครงสร้างของระบบการแปลภาษาด้วยคอมพิวเตอร์โดยทั่วไป

สำหรับระบบการแปลภาษาด้วยคอมพิวเตอร์ของงานวิจัยนี้ ใช้กระบวนการแปลภาษาโดยมีภาษากลาง (Interlingua) เป็นตัวเชื่อมระหว่างภาษาต้นฉบับ (Source Language) กับภาษาเป้าหมาย (Target Language) ดังแสดงในรูปที่ 2 ซึ่งมีกระบวนการแบ่งเป็น 2 ขั้นตอนใหญ่ๆ นั่นคือขั้นตอนการวิเคราะห์ภาษาต้นฉบับไปสู่ภาษากลาง (ด้านซ้ายของรูปที่ 2) และขั้นตอนการสร้างภาษาเป้าหมายจากภาษากลาง (ได้แก่ด้านขวาของรูปที่ 2) วิธีการแปลภาษาแบบนี้จะช่วยลดความยุ่งยากและลดเวลาในการวิเคราะห์ภาษาต้นฉบับ และ การสร้างภาษาเป้าหมาย เมื่อเปรียบเทียบกับกระบวนการแปลภาษาแบบถ่ายทอดโดยตรง (Direct MT Strategy)[1]

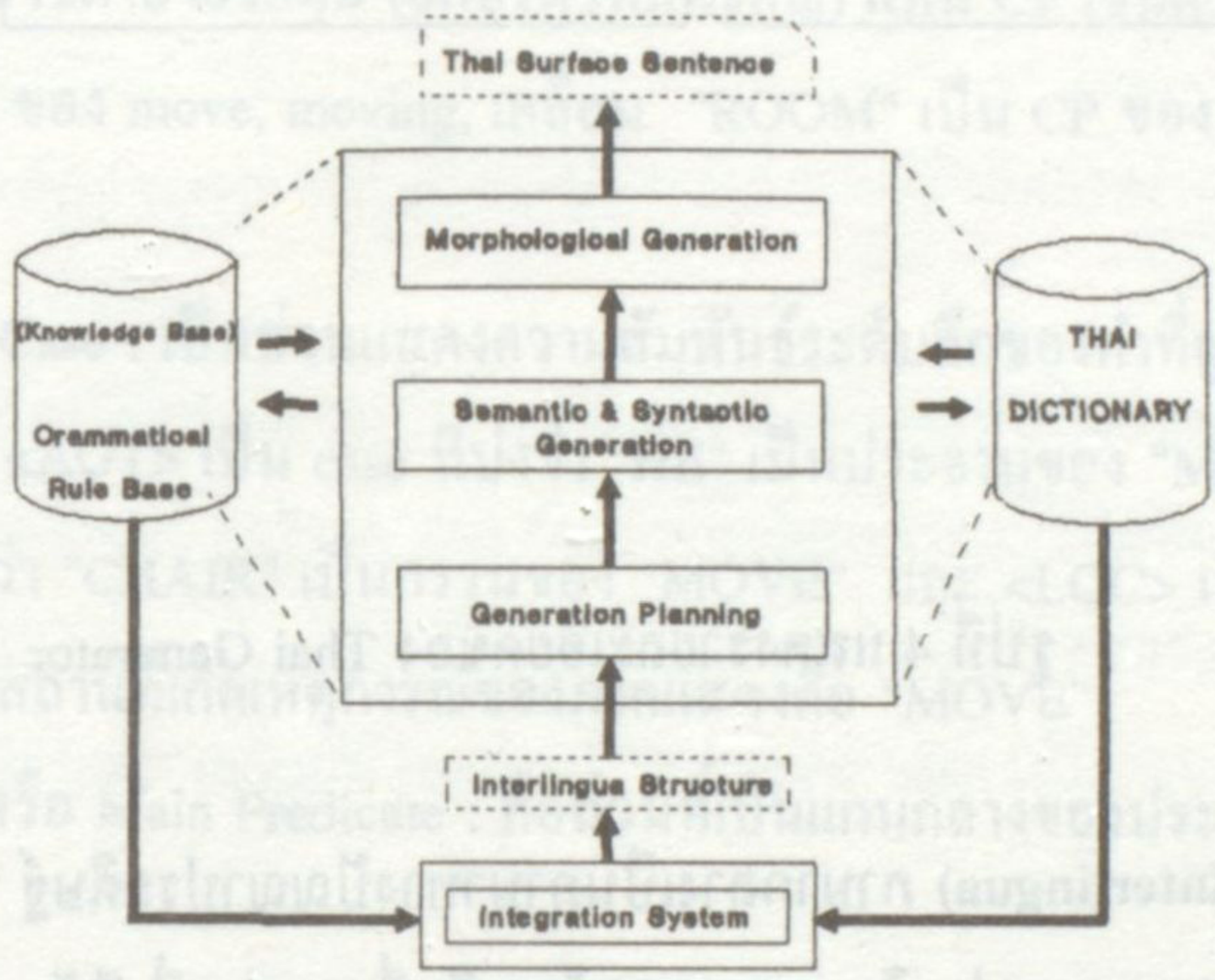



รูปที่ 2 โครงสร้างพื้นฐานของระบบแปลภาษาของงานวิจัยนี้

ทั้งนี้เพราะเราไม่จำเป็นต้องคำนึงถึงคู่ภาษาที่จะแปล เนื่องจากการวิเคราะห์ภาษาต้นฉบับใดๆ ก็ตามให้อยู่ในรูปของภาษากลางแล้วก็จะสามารถแปลเป็นภาษาเป้าหมายอะไรก็ได้ โดยที่ไม่ต้องวิเคราะห์ใหม่ และในทำนองเดียวกัน ทางด้านการสร้างภาษาเป้าหมายใดๆ นั้น ก็เป็นการสังเคราะห์จากภาษากลาง ซึ่งเป็นผลของการวิเคราะห์ภาษาต้นฉบับใด ๆ เพียงอย่างเดียว เมื่อได้มีการสังเคราะห์แล้วก็ไม่ต้องทำซ้ำอีกต่อไป นอกจากนั้นการแปลภาษาแบบถ่ายทอดทางอ้อมโดยใช้ภาษากลางนี้จะใช้พจนานุกรมภาษากลางเป็นฐานข้อมูลในการวิเคราะห์ และสังเคราะห์เพียงชุดเดียวต่อหนึ่งภาษา ซึ่งต่างจากการแปลภาษาแบบถ่ายทอดโดยตรง ที่จะต้องใช้พจนานุกรมถึง n ชุด ในการแปลจากภาษาต้นฉบับ 1 ภาษาไปสู่ภาษาเป้าหมาย n ภาษา

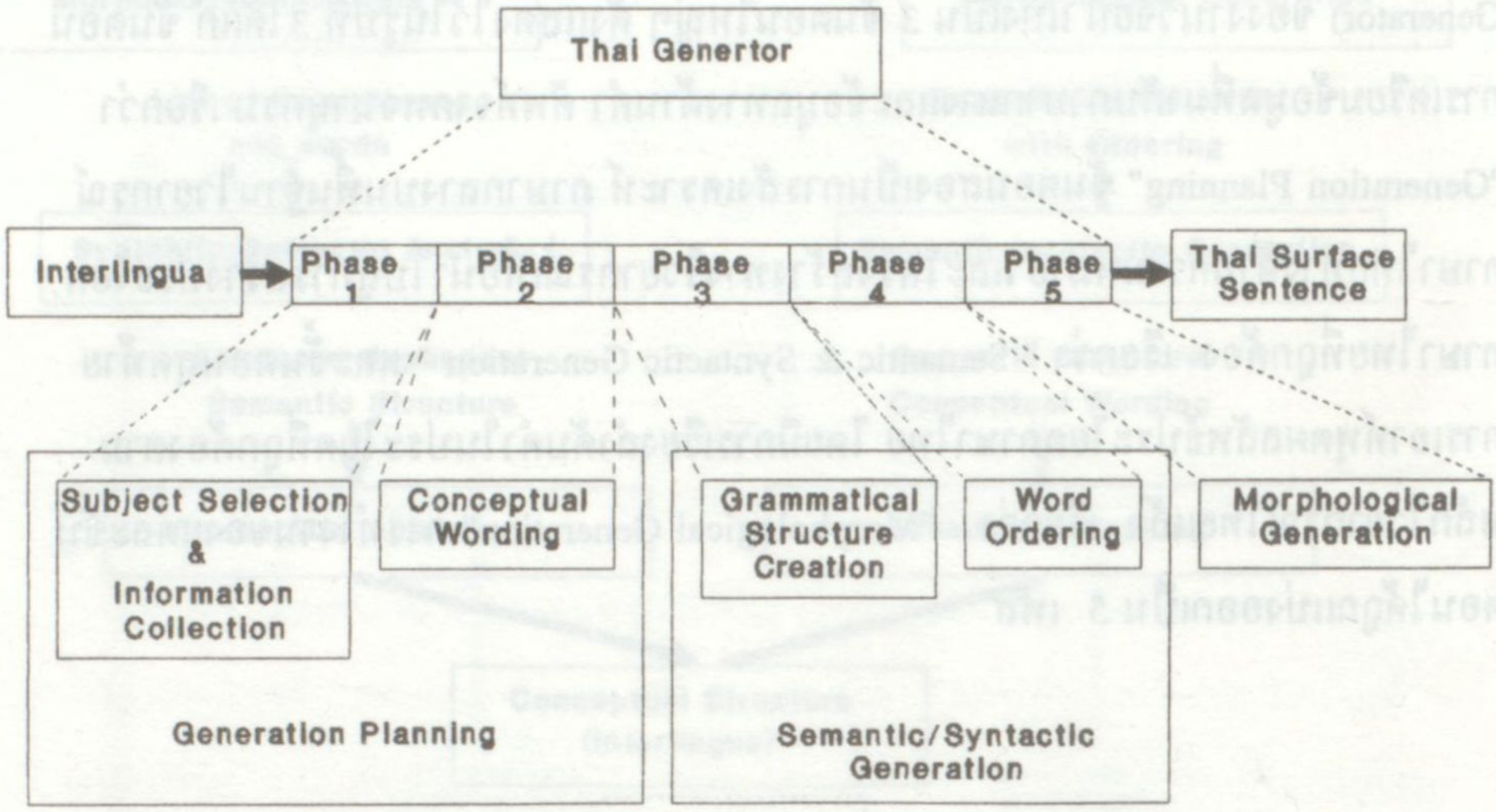
ในการสร้างประโยคภาษาไทย จากภาษากลางของงานวิจัยนี้ได้ทำการศึกษาถึงโครงสร้างของภาษากลาง ซึ่งจะต้องถูกกำหนดขึ้นให้เป็นตัวแทนทางด้านความหมายลึกของทุกภาษา โดยอาศัยทฤษฎีทางวากยสัมพันธ์ หรือไวยากรณ์เคส (Case Grammer) [2] ของฟิลล์มอร์เป็นพื้นฐานและพิจารณาให้เหมาะสมกับภาษาไทย เพื่อจะได้นำไปสู่การเขียน

กระบวนการสร้างภาษาไทยจากภาษากลางดังกล่าวกระบวนการ สร้างภาษาไทย (Thai Generator) ของงานวิจัยนี้ แบ่งเป็น 3 ขั้นตอนใหญ่ๆ ดังแสดงไว้ในรูปที่ 3ได้แก่ ขั้นตอนการเตรียมข้อมูลที่มากับภาษากลางและข้อมูลทางด้านคำ ศัพท์จากพจนานุกรม เรียกว่า "Generation Planning" ขั้นตอนสองเป็นการสังเคราะห์ ภาษากลางบนพื้นฐานไวยากรณ์ภาษาไทยทางด้านความหมาย และโครงสร้างทางไวยากรณ์เพื่อนำไปสู่การสร้างประโยคภาษาไทยที่ถูกต้อง เรียกว่า " Semantic & Syntactic Generation " และขั้นตอนสุดท้ายการเอาท์พุตผลลัพธ์ประโยคภาษาไทย โดยมีการเรียงลำดับคำในประโยคที่ถูกต้องตามหลักไวยากรณ์ไทยแล้ว เรียกว่า "Morphological Generation" การทำงานของแต่ละขั้นตอนได้ถูกแบ่งออกเป็น 5 เฟส




รูปที่ 3 โครงสร้างของระบบการสร้างภาษาไทยจากภาษากลางของงานวิจัยนี้

เพื่อให้ง่ายต่อการพัฒนาตัว Generator ซึ่งเป็นซอฟท์แวร์และการแก้ไขปรับปรุงในภายหลัง ความสัมพันธ์ในด้านหน้าที่ของแต่ละ Phase ได้แสดงไว้ในรูปที่ 4




รูปที่ 4 แสดงรายละเอียดของ Thai Generator

**2. ภาษากลาง (Interlingua)** ภาษากลางเป็นภาษาทางปัญญาประดิษฐ์ ที่ถูกกำหนดขึ้นเพื่อแสดงความหมายของประโยคของภาษาต้นฉบับที่จะแปล ซึ่งมีลักษณะเป็นกลางไม่ขึ้นกับภาษาใด (Language Independent) และสามารถใช้แทนทุกภาษา (Multilanguage Representation) การวิจัยทางด้านการกำหนดรูปแบบของภาษากลาง เพื่อแทนความหมายของภาษาต่างๆ นั้นมีอยู่ในระบบแปลภาษาหลายระบบ[1] ซึ่งโดยส่วนใหญ่มีการกำหนดโครงสร้างของภาษากลางเป็น 2 ส่วน ได้แก่ส่วนแสดงหน่วยความหมาย (Unit of meaning) ของคำในภาษาต้นฉบับ เรียกว่า Conceptual Primitive (CP) และส่วนแสดงความสัมพันธ์ของหน่วยความหมาย เพื่อแทนความหมายของเนื้อหาในภาษาต้นฉบับ เรียกส่วนนี้ว่า ไวยากรณ์ภาษากลาง(Interlingual Grammar) ในการกำหนด CP เพื่อแทนคำในแต่

ละภาษา จะต้องพิจารณาความหมายลึก (deep meaning หรือ concept) ของคำนั้น ประกอบกับคุณสมบัติ (property) ของคำนั้นๆ ด้วย โดยการจัด หมวดหมู่ของคำไปตามข้อกำหนดทั้งสองสำหรับภาษากลางที่ใช้ในงานวิจัยนี้ ก็มีลักษณะดังกล่าวข้างต้น โดยในส่วนการแสดงความสัมพันธ์ของ CP นั้นอาศัย ทฤษฎีทางการกสัมพันธ์ (Case Relation) และแสดงโครงสร้างภาษากลาง ในรูปความ สัมพันธ์พึ่งพิง (Conceptual Structure หรือ Dependency Tree Representation) ส่วนประกอบภาษากลางที่เป็นอินพุตของระบบการสร้างภาษาไทยของงานวิจัยนี้มี

1) Conceptual Primitive (CP) : เป็นส่วนที่ใช้แทนคำในแต่ละภาษาที่มีความหมายลึก และลักษณะการใช้ในประโยคเหมือนกัน โดยมีการจัดกลุ่มคำเหล่านั้นไว้ใน concept เดียวกัน เรากำหนดสัญญลักษณ์สำหรับแทน "CP" ของกลุ่มคำในแต่ละภาษา ดังกล่าวเช่นไรก็ได้ ในงานวิจัยนี้จะใช้ภาษาอังกฤษ (อักษรตัวพิมพ์ใหญ่) แทน CP เช่นตัวอย่างในรูปที่ 5 "MOVE" เป็น CP ของ move, moving, เคลื่อน "ROOM" เป็น CP ของ room, ห้อง เป็นต้น

2) การก หรือ Case : เป็นส่วนแสดงความสัมพันธ์ระดับลึกของคำที่ถูกแทน ด้วย CP เช่นตัวอย่างรูปที่ 5 <AGT> เป็น case ที่บ่งว่า "HE" เป็นประธานของ "MOVE" และ <OBJ> เป็น case ที่บ่งว่า "CHAIR" เป็นกรรมของ "MOVE" และ <LOC> เป็น case ที่บ่ง ว่า "ROOM" เป็นสถานที่เกิดเหตุการณ์ของภาคแสดงคือ "MOVE"

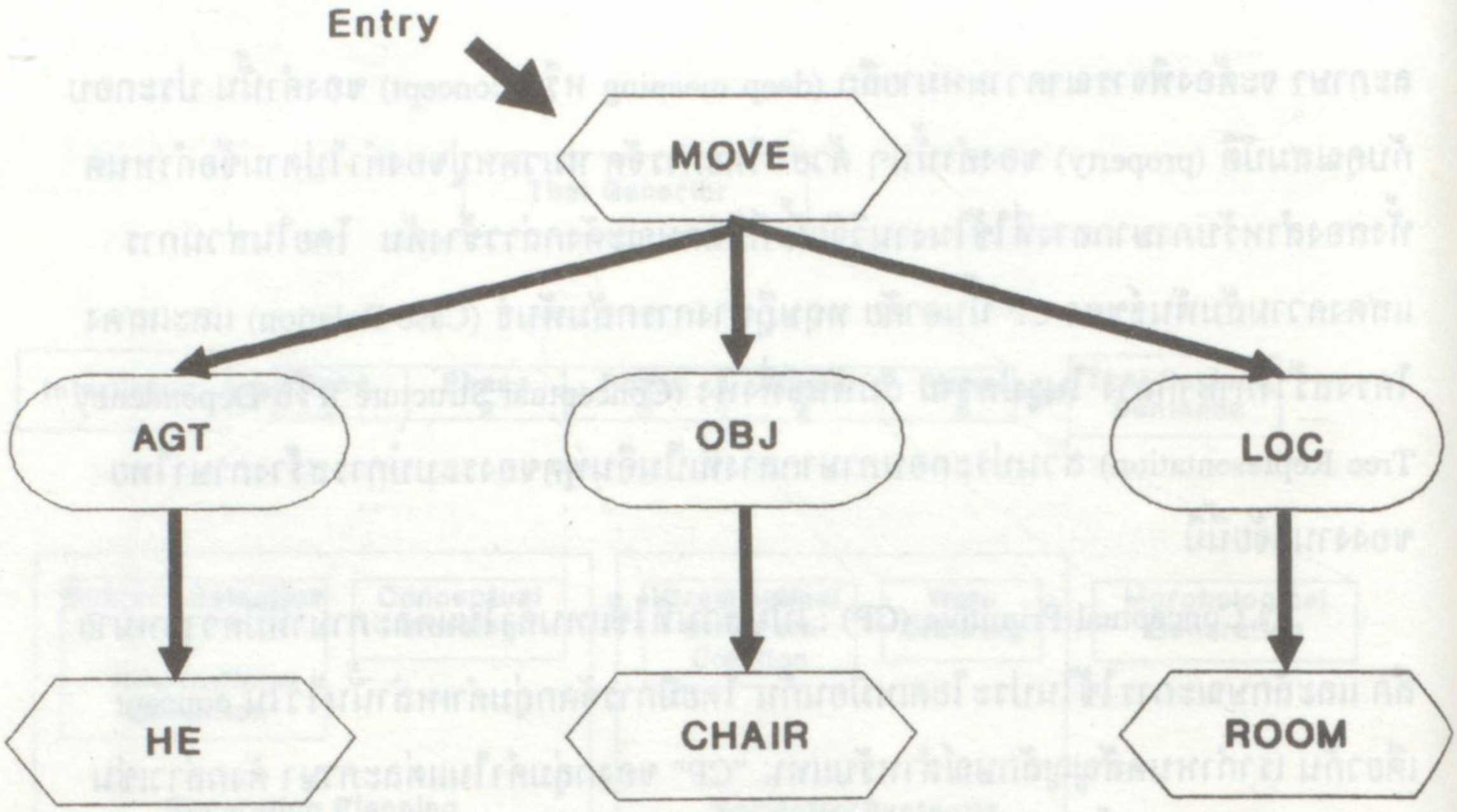
3) กริยาหลัก หรือ Main Predicate : คือส่วนที่เป็นแกนกลางของประโยค โดยจะพิจารณาจาก

ก) คำกริยา เช่น เคลื่อน กิน วิ่ง เดิน นั่ง อ่าน ร้อง

ข) คำคุณศัพท์ที่ทำหน้าที่คล้ายคำกริยา เช่น สวย หอม ใหญ่ สูง

ค) คำกริยาช่วย เช่น เป็น อยู่ คือ verb to be, there be, have

ในโครงสร้างของภาษากลางนี้ กริยาหลักจะเป็นจุด หรือโนดเริ่มต้น (entry node) ของการวิเคราะห์และสร้างประโยค ตัวอย่างรูปที่ 5 คือ "MOVE"




รูปที่ 5 โครงสร้างภาษากลางของประโยค "He moves a chair in the room."

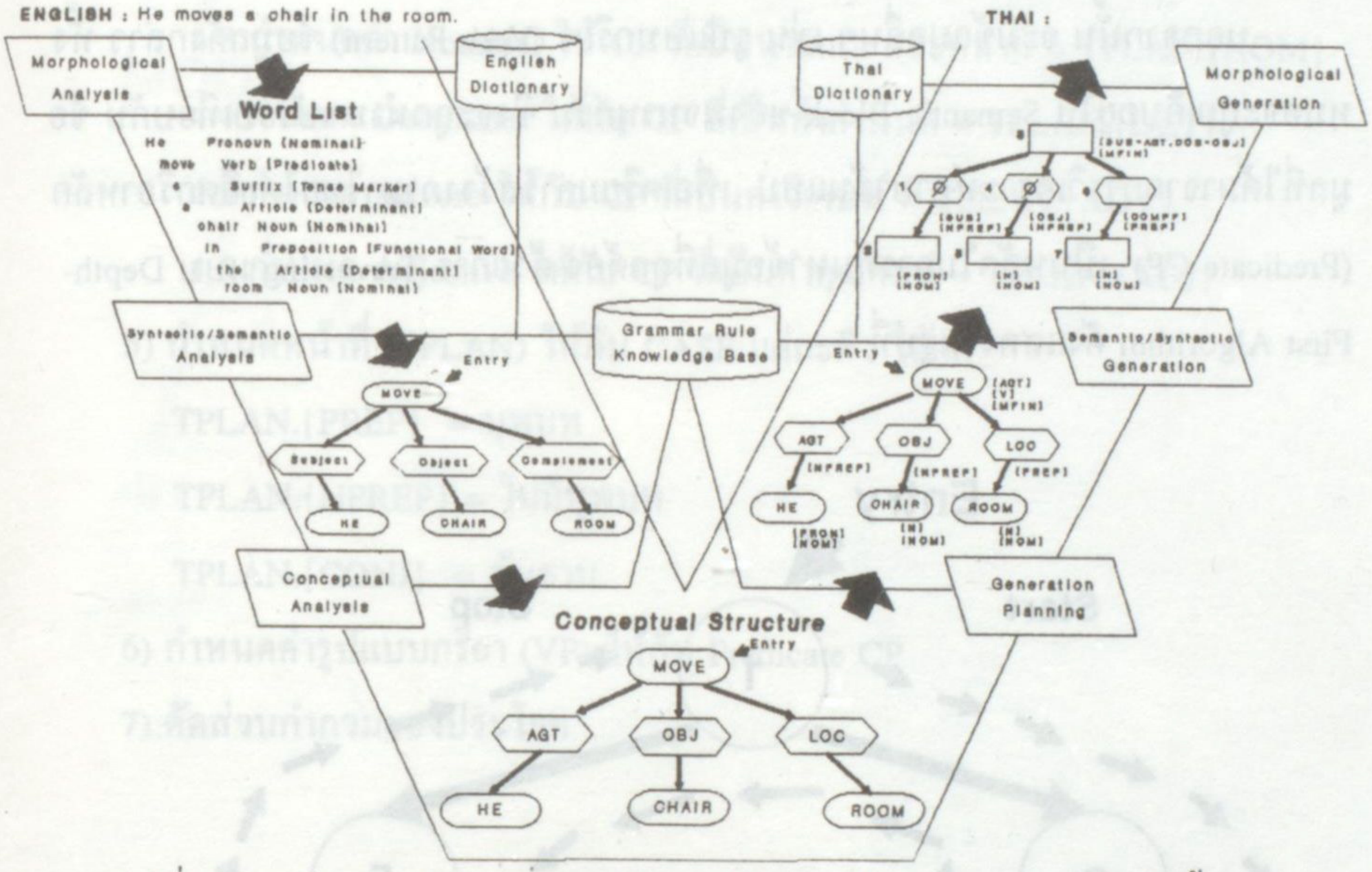
4) Dependency Tree Representation : เป็นส่วนแสดงโครงสร้างของภาษากลางในรูปแบบของต้นไม้ ประกอบด้วย "โนด" และ "อาร์ค" โนดจะแทน CP และ Case ส่วนอาร์คแสดงความสัมพันธ์ของโนดในลักษณะ head-dependent relation ดังตัวอย่างรูปที่ 5 "MOVE" เป็น head ของ "HE" ในทางตรงกันข้าม "HE" เป็น dependent ของ "MOVE"

5) Feature : เป็นส่วนแสดงข้อมูลทางไวยากรณ์ของ CP และ Case ที่ได้จากการวิเคราะห์ภาษาต้นฉบับ

**3. ระบบการสร้างภาษาไทยจากภาษากลาง (Thai Generation System from Interlingua)** การสร้างประโยคภาษาไทยจากภาษากลาง จะอาศัยข้อมูลที่ได้จากฝ่ายวิเคราะห์ภาษาต้นแบบซึ่งเก็บไว้ในภาษากลางเป็น Feature ของ CP และ Case พร้อมกับข้อมูลของคำเหล่านั้นจากพจนานุกรมภาษาไทย ระบบจะต้องค้นหาคำในภาษาไทยที่ทำหน้าที่ และมีความหมายตรงกับ CP และCase ในภาษากลาง และจะต้องตัดคำกำกวมโดยเลือกเอา

เฉพาะคำที่ถูกต้องที่สุด ในขั้นตอนสุดท้ายจะต้องทำการเรียงคำให้อยู่ ในตำแหน่งที่ถูกต้อง พร้อมกับมีการตกแต่งประโยคให้ถูกต้องและสมบูรณ์ ตามหลักไวยากรณ์ภาษาไทย ดังตัวอย่างที่แสดงในรูปที่ 6

โครงสร้างของระบบการสร้างภาษาไทยในงานวิจัยนี้ แบ่งออกเป็น 3 ขั้นตอนได้แก่การเตรียมข้อมูลสำหรับการสร้างประโยคภาษาไทย(Generation Planning) การสร้างโครงสร้างสำหรับประโยคตามไวยากรณ์ภาษาไทย (Semantic and Syntactic Generation) และการกำหนดคำภาษาไทย (Morphological Generation) ดังแสดงในรูปที่ 3



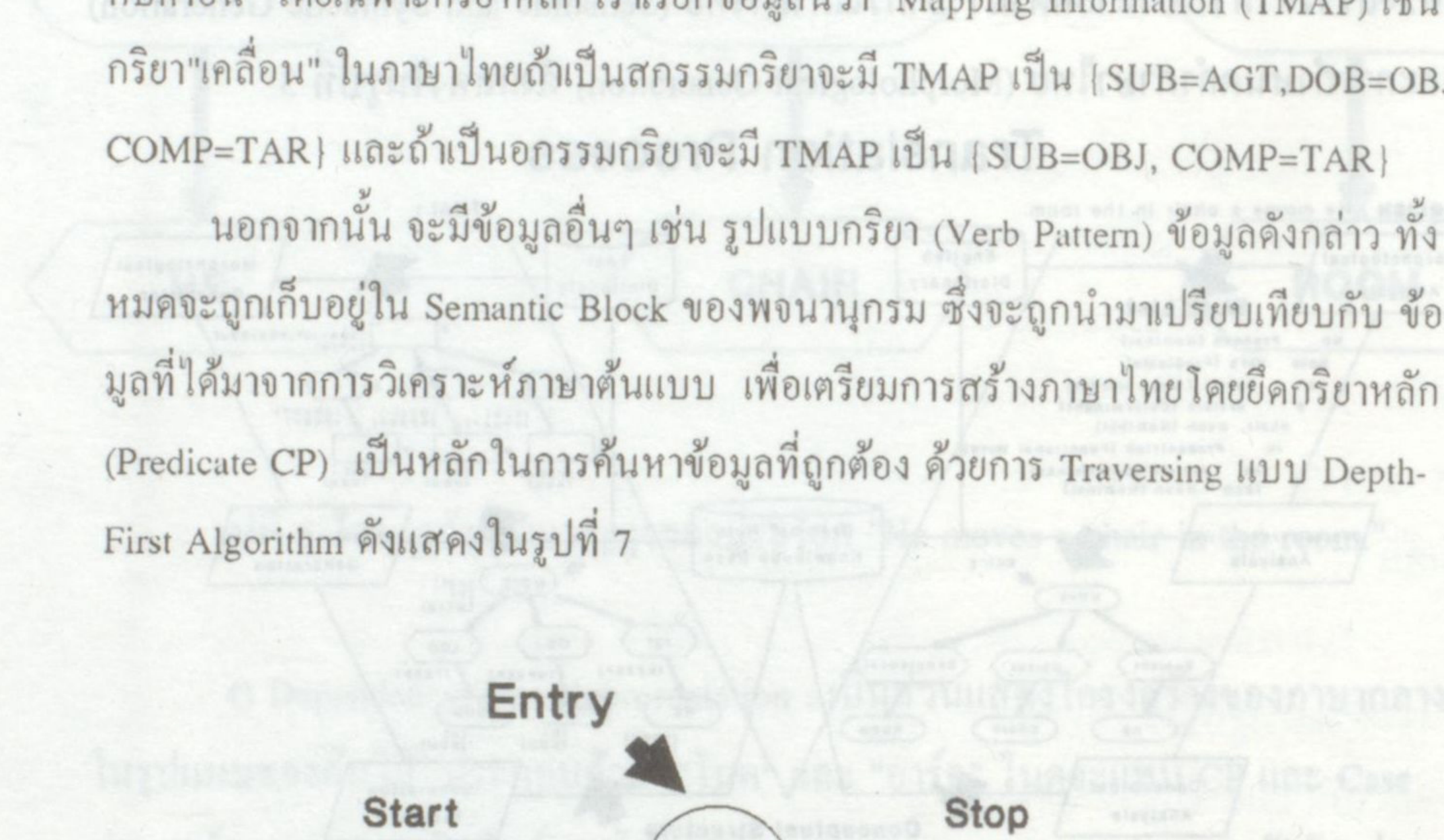

รูปที่ 6 ตัวอย่างการแปลจากภาษาอังก-ษเป็นภาษาไทย แสดงให้เห็นถึงขั้นตอนการทำงานทั้งหมดของระบบการแปลภาษาด้วยคอมพิวเตอร์ของงานวิจัยนี้

**3.1 Generation Planning** ในส่วนนี้จะทำการดึงข้อมูลจากพจนานุกรมภาษาไทย โดยใช้ชื่อ CP หรือ Case จากภาษากลางเป็น key word หรือตัวกำหนดตำแหน่งของข้อมูลจากพจนานุกรม ข้อมูลจะประกอบด้วยคำในภาษาไทย และคุณสมบัติทางด้านไวยากรณ์ของคำนั้นๆ ได้แก่

ก) คุณสมบัติทางด้าน Thai Category (TCAT) ของคำ ได้แก่นาม (N) สรรพนาม (PRON) กริยา (V) คุณศัพท์ (ADJ) กริยาวิเศษณ์ (ADV) บุพบท (PREP) สันธาน (CONJ) เป็นต้น

ข) คุณสมบัติทางด้านโครงสร้างที่บ่งความสัมพันธ์ระดับลึกและระดับผิวของคำนั้นๆ กับคำอื่น โดยเฉพาะกริยาหลัก เราเรียกข้อมูลนี้ว่า Mapping Information (TMAP) เช่น กริยา"เคลื่อน" ในภาษาไทยถ้าเป็นสกรรมกริยาจะมี TMAP เป็น {SUB=AGT,DOB=OBJ, COMP=TAR} และถ้าเป็นอกรรมกริยาจะมี TMAP เป็น {SUB=OBJ, COMP=TAR}

นอกจากนั้น จะมีข้อมูลอื่นๆ เช่น รูปแบบกริยา (Verb Pattern) ข้อมูลดังกล่าว ทั้งหมดจะถูกเก็บอยู่ใน Semantic Block ของพจนานุกรม ซึ่งจะถูกนำมาเปรียบเทียบกับ ข้อมูลที่ได้มาจากการวิเคราะห์ภาษาต้นแบบ เพื่อเตรียมการสร้างภาษาไทยโดยยึดกริยาหลัก (Predicate CP) เป็นหลักในการค้นหาข้อมูลที่ถูกต้อง ด้วยการ Traversing แบบ Depth-First Algorithm ดังแสดงในรูปที่ 7




รูปที่ 7 หลักการ traversing แบบ depth-first algorithm

หน้าที่หลักในส่วนของ Generation Planning มีตามลำดับดังต่อไปนี้คือ

1) เปลี่ยนโครงสร้างของ Conceptual Structure ในกรณีจำเป็นได้แก่การเพิ่มโนดใหม่ที่เป็น CP และ Case

2) กำหนดประธานของประโยคให้กับกริยาหลัก (Predicate CP)

3) กำหนดค่า TCAT ให้กับ CP แต่ละตัว

4) กำหนดหน้าที่ (TPLAN) ให้กับ CP แต่ละตัว ดังนี้

4.1 กำหนดค่า Main Finite ให้กับ predicate CP = TPLAN.{MFIN}

4.2 กำหนดค่า Finite ให้กับกริยาย่อยตัวอื่น = TPLAN.{FIN}

4.3 กำหนดค่า Nominal ให้ CP ที่เป็นนามและสรรพนาม = TPLAN.{NOM}

4.4 กำหนด Deteminer ให้กับ CP ที่เป็นค่ากำหนด = TPLAN.{DET}

4.5 กำหนด Adverb ให้กับ CP ที่เป็นคำวิเศษณ์ = TPLAN.{ADV}

4.6 กำหนด Adjective ให้กับ CP ที่เป็นคำคุณศัพท์ = TPLAN.{ADJ}

5) กำหนดหน้าที่ (TPLAN) ให้กับ CASE แต่ละตัว ได้แก่

TPLAN.{PREP} = บุพบท

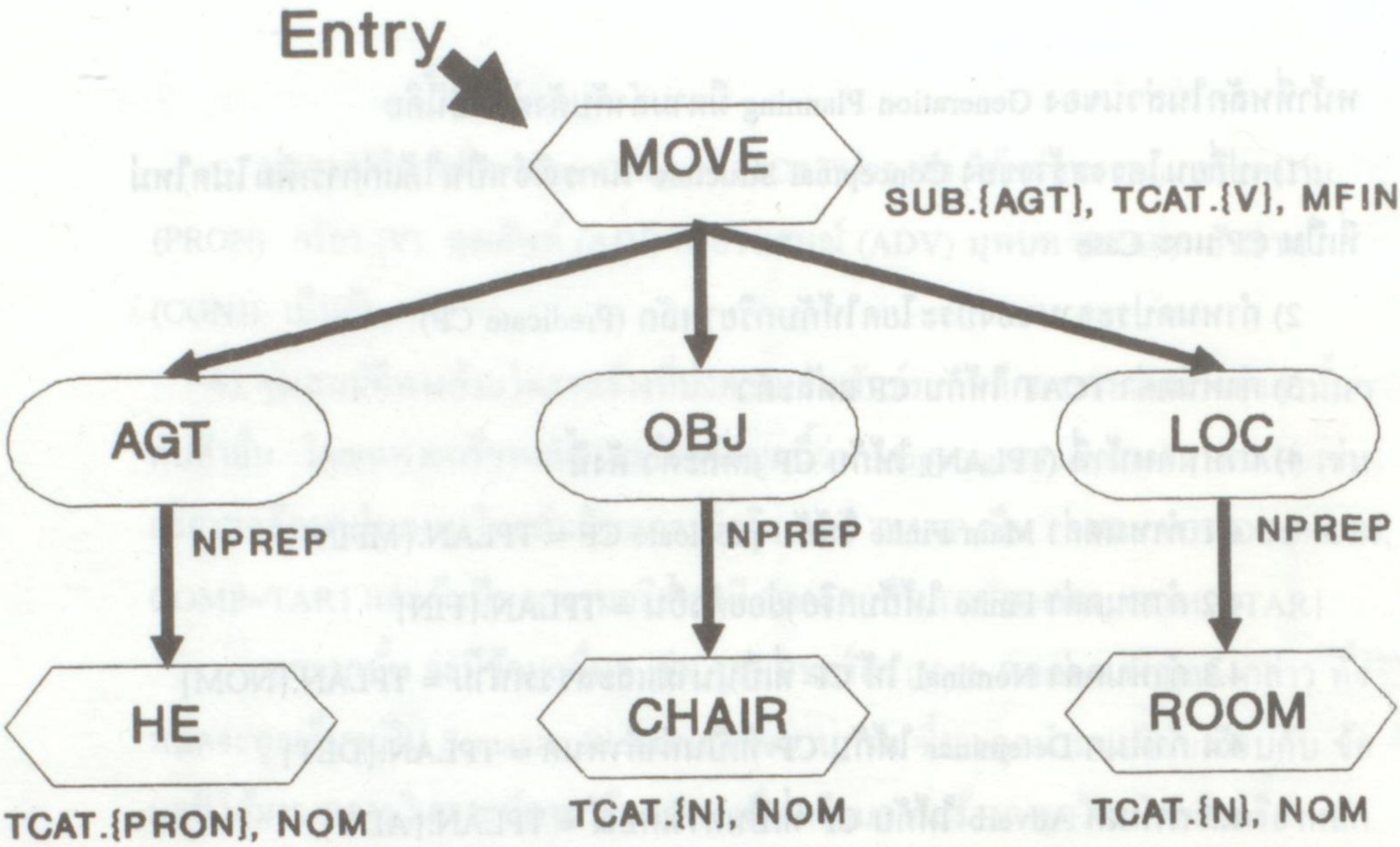
TPLAN.{NPREP} = ไม่มีบุพบท

TPLAN.{CONJ} = สันธาน

6) กำหนดค่ารูปแบบกริยา (VP) ให้กับ Predicate CP

7) ตัดส่วนกำกวมของประโยค

ในบทความวิจัยนี้ จะใช้ประโยคภาษาอังก-ษ "He moves a chair in the room." เป็นตัวอย่างที่จะแปลเป็นภาษาไทย ประโยคนี้จะมีรูปแบบภาษากลางดังแสดง ในรูปที่ 5 เมื่อภาษากลางนี้ผ่านส่วน Generation Planning จะได้ผลลัพธ์สุดท้ายของประโยค นี้ดังแสดงไว้ในรูปที่ 8




รูปที่ 8 ผลลัพธ์ที่เป็นส่วนเอาท์พุตของ Generation Planning
สำหรับ การแปลประโยค "He moves in the room."

**3.2 Semantic and Syntactic Generation** การทำงานของขั้นตอนนี้จะถูกแบ่งเป็น

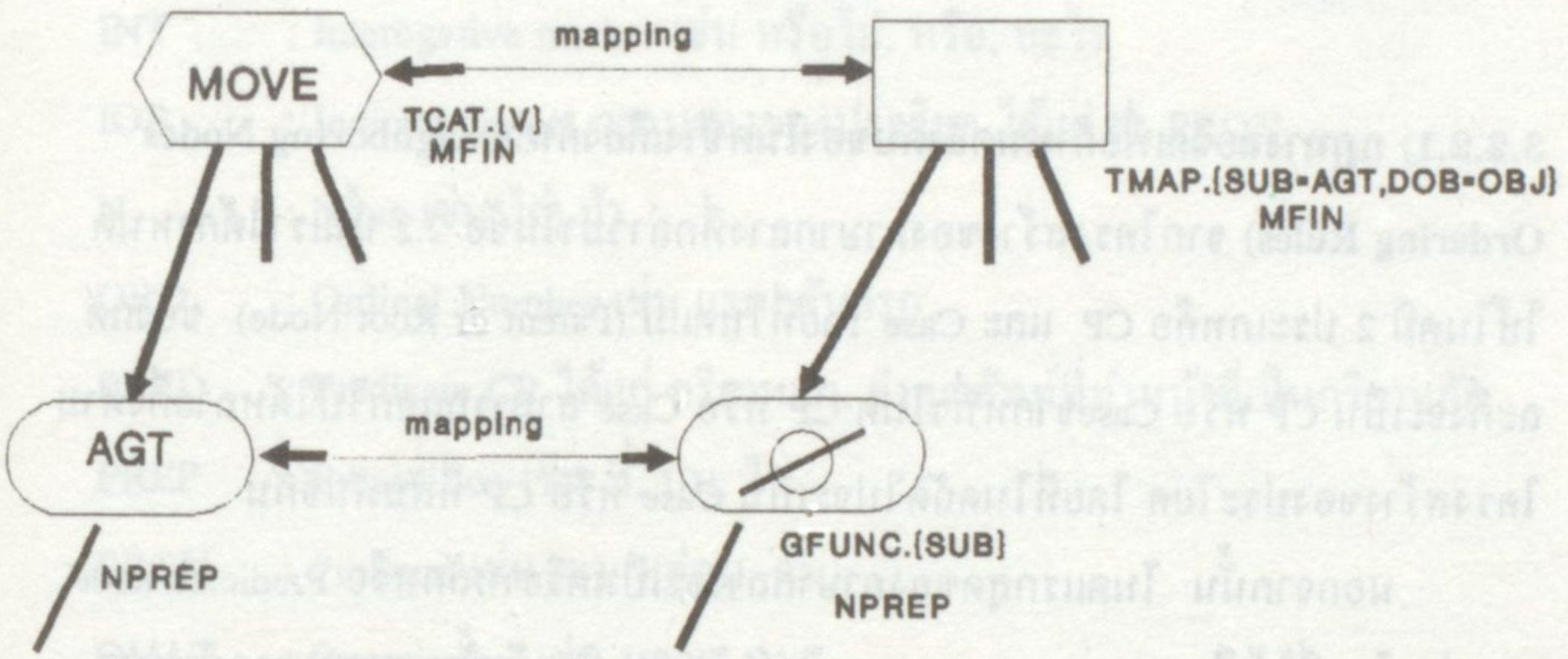
**3.2.1 การสร้างโครงสร้างประโยคทางไวยากรณ์ (Grammatical Structure Creation)** ในส่วนนี้จะทำการสร้างโครงสร้างทางไวยากรณ์ของภาษาไทย (Grammatical Structure) ให้กับภาษากลาง โดยทำการกำหนดหน้าที่ของคำในประโยค ของ CP ในฟิลด์ชื่อ GFUNC แต่ละตัวที่เชื่อมต่อจากกริยาหลักตามหลักไวยากรณ์ไทยและใส่ค่านั้นไว้ที่ Case ที่แสดงความสัมพันธ์ระหว่างกริยาหลักกับ CP ตัวนั้นๆ ได้แก่

1. GFUNC.{SUB} สำหรับ CP ที่เป็นประธาน (Subject)
2. GFUNC.{DOB} สำหรับ CP ที่เป็นกรรมตรง (Direct Object)
3. GFUNC.{IOB} สำหรับ CP ที่เป็นกรรมรอง (Indirect Object)

4. GFUNC.{COMP} สำหรับ CP ที่เป็นส่วนขยายประโยคที่ขาดไม่ได้

5. GFUNC.{COMPF} สำหรับ CP ที่เป็นส่วนขยายประโยคทั่วไป

รูปที่ 9 แสดงการกำหนด Grammatical Structure ให้กับประโยคตัวอย่างโดยใช้วิธีการเทียบ (Mapping) Conceptual CP เข้ากับ Grammatical Node ซึ่งได้ผลลัพท์สำหรับตัวอย่างนี้ในรูปที่ 10 หลังจากนั้นระบบจะทำการเรียงลำดับของคำที่ถูกต้องตามหลักไวยากรณ์ไทยในขั้นตอนต่อไป




รูปที่ 9 ตัวอย่างการกำหนดโครงสร้างทางไวยากรณ์

**3.2.2 การเรียงลำดับของคำภาษาไทย (Thai Word Ordering)** การเรียงลำดับของคำภาษาไทยที่อยู่บนโครงสร้างไวยากรณ์ในข้อ 3.2.1 นั้น ก็คือการจัดลำดับของโนด CP และโนด Case เพื่อที่จะนำมาเรียงกันให้เป็นประโยคในแนวเส้นตรง ทั้งนี้โดยอาศัยข้อมูลที่กำหนดตำแหน่งของคำที่อยู่ข้างเคียงกันตามหลักโครงสร้าง (Syntax) ของภาษา

ไทยมาเป็นตัวตัดสิน ดังนั้นจึงต้องมีการกำหนดกฎการเรียงลำดับคำหรือกลุ่มคำ โดยแบ่งตามประเภทของคำ และหน้าที่ของคำนั้น และใช้กฎดังกล่าวเป็นฐานข้อมูลในการเปรียบเทียบอ้างอิงของการกำหนดลำดับของโนด CP และโนด Case ที่อยู่ในโครงสร้างต้นไม้ เพราะฉะนั้น การทำงานในส่วนนี้ จะประกอบไปด้วยส่วนสำคัญ 2 ส่วนคือ กฎทางลอจิกเพื่อกำหนดลำดับของโนดที่อยู่ข้างเคียงกันบโครงสร้าง ต้นไม้(ในที่นี้คือโครงสร้างภาษากลางหรือ Conceptual Structure) และ ส่วนที่เป็นฐานข้อมูล ของกฎไวยากรณ์การเรียงลำดับคำและกลุ่มคำในภาษาไทย ซึ่งจะกล่าวโดยละเอียดต่อไปผลลัพธ์ของตัวอย่างประโยคที่แสดงในรูปที่ 10 นั้น ตัวเลขที่กำกับอยู่กับโนดต่างๆ ทั้ง CP และ CASE คือ ตัวเลขที่บอกลำดับการเรียงคำในประโยคภาษาไทย

**3.2.2.1) กฎทางลอจิกเพื่อกำหนดลำดับของโนดข้างเคียงกัน (Neighboring Nodes Ordering Rules)** จากโครงสร้างของภาษากลางที่กล่าวมาในข้อ 2.2 นั้นเราได้กำหนดให้โนดมี 2 ประเภทคือ CP และ Case โดยที่โนดแม่ (Parent or Root Node) ของแต่ละกิ่งจะเป็น CP หรือ Caseจากหนึ่งโนด CP หรือ Case สามารถแตกไปได้หลายกิ่งตามโครงสร้างของประโยค โดยที่โนดถัดไปจะเป็น Case หรือ CP ที่สัมพันธ์กัน

นอกจากนั้น โนดแรกสุดของภาษากลางจะเป็นกริยาหลักหรือ Predicate node ของประโยคที่ใช้เป็น entry node ของการจัดลำดับของคำ ดังนั้น เราสามารถกำหนดลำดับของโนดด้วยการเปรียบเทียบความสัมพันธ์ของโนดเป็นคู่ ๆ ที่อยู่ข้างเคียงกันโดยเริ่มจาก entry node ซึ่งเป็น root node ตัวแรกสุดไปตามกิ่งจากบนลงล่าง และจากซ้ายไปขวา(ตามหลักการDepth-First Algorithm) ตามลำดับจนกระทั่งย้อนกลับมาที่ entry node อีกครั้งหนึ่ง กฎทั้งหมดแบ่งได้เป็น 2 กรณีคือ

(1) การจัดเรียงลำดับในแนวดิ่ง (Head - Dependent Ordering) จะเป็นการตรวจสอบในแนวดิ่งว่า CP ใดมาก่อนหรืออยู่หลัง CP ใดในกิ่งเดียวกัน ซึ่งการจัดลำดับจะใช้ตารางที่ 1 เป็นกฎการตรวจสอบ นอกจากนั้นีจะทำการตรวจสอบอีกว่า กิ่งที่ต่อมาจาก Predicate node จะอยู่ข้างหน้าหรืออยู่ข้างหลัง Predicate node ที่ต่ออยู่

ความหมายของสัญญลักษณ์ในตารางที่ 1 มีดังต่อไปนี้

ADJ : Adjective เช่น สวย, ใหญ่, เล็ก

CAD : Cardinal Number เช่น 1, 2, สาม, ห้า, ร้อย

CLAS : Classifier เช่น คัน, ตัว, เล่ม, ด้าม

COMP : Complement เป็นส่วนที่อยู่หลัง IOB

COMPF : Complement ที่บอกเวลา, บอกเหตุผล

DEM : Memonsative เช่น นี่, นั่น

DOB : Direct Object กรรมตรงของประโยค ได้แก่ N, PRON

INT : Interogative marker เช่น หรือไม่, หรือ, อะไร

IOB : Indirect Object กรรมรองของประโยค ได้แก่ N, PRON

N : Noun เช่น ไก่, ม้า

ORD : Ordinal Number เช่น แรก, คันแรก

PRED : Predicate CP ได้แก่ กริยาหลัก, คำคุณศัพท์ที่ทำหน้าที่เป็นกริยาหลัก

PREP : Preposition เช่น ที่, บน, ใน

PRON : Pronoun เช่น เขา, หล่อน, ฉัน

QUAT : Quantitative เช่น หลาย, บาง

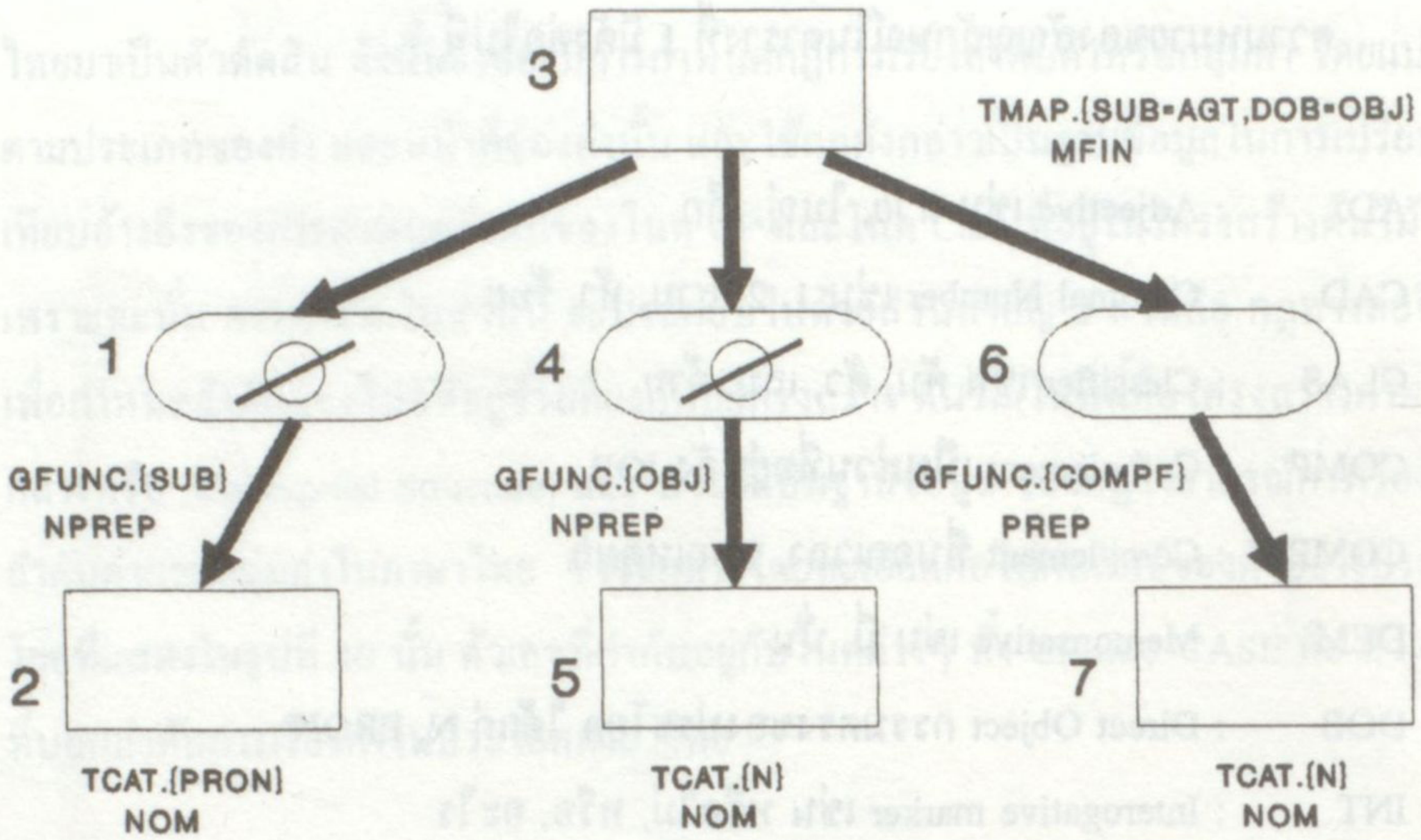
SADV : Sentencial Adverb ส่วนขยายความประโยค เช่น โดยทั่วไป, ที่จริง, ตามความจริง, ดังนั้น

SMOD : Sentencial Modefier ส่วนขยายประโยค เช่น เมื่อเร็วๆนี้, กาลครั้งหนึ่ง

SUB : Subject ประธานของประโยค

V : Verb เช่น กิน, เดิน, นั่ง

(2) การจัดเรียงคำในแนวนอน(Sister-Sister Ordering) เป็นการตรวจสอบ CP และ Case node ในแนวนอน โดยจะตรวจสอบจาก CP ที่เป็น Predicate nodeว่า Case ที่ต่ออยู่นั้นมี ค่า GFUNC เป็นอะไร แล้วให้คะแนนเก็บไว้ที่ Case ซึ่ง Case ใดมีคะแนนน้อยจะมาก่อน Case ที่มีคะแนนมากซึ่งการให้คะแนนแก่ Case




รูปที่ 10 ผลลัพธ์ของโครงสร้างทางไวยากรณ์ภาษาไทย สำหรับการแปลประโยค

"He moves a chair in the room.

ตารางที่ 1 เป็นตารางการจัดเรียงลำดับ CP ในแนวดิ่ง

```
SADV < PRED
SMOD < PRED
SUB  < PRED
PRED < DOB, IOB, COMP, COMPF
PRED < INT
PREP < N, PRON
N1   < V, CLAS, PREP, N2, PRON
N1   < PRON, N3
CLAS < ORD, DEM, ADJ
N2   < CAD, QUANT, CLAS
N2   < CAD, QUANT
N1,2 < CLAS
CAD  < CLAS
QUANT < CLAS
```

หมายเหตุ เครื่องหมาย A < B หมายถึง A มีลำดับอยู่หน้า B

จะให้ตามตารางที่ 2

ตารางที่ 2 แสดงการให้แต้มแก่ CASE ในการจัดลำดับในแนวนอน

| MONITOR | FIELD | DATA | VALUE |
|---|---|---|---|
| | | | |
| SMOD | CASE | SMOD | 1000 |
| SUB | GFUNC | SUB | 2000 |
| NEG | NEG | T | 3000 |
| DOB | GFUNC | DOB | 4000 |
| IOB | GFUNC | IOB | 5000 |
| TAR | CASE | TAR | 6000 |
| COMP | GFUNC | COMP | 7000 |
| COMPF | GFUNC | COMPF | 8000 |
| END | ESUBCAT | END | 9000 |

**4.บทสรุป** ระบบการสร้างภาษาเป้าหมายเพื่อการแปลภาษาของงานวิจัยนี้ เป็นการวิจัยและพัฒนาระบบการสร้างภาษาไทยจากภาษากลาง ซอฟท์แวร์ของระบบการแปลเขียนด้วยภาษาC ส่วนกฎไวยากรณ์สำหรับการสร้างภาษาไทยเขียนด้วยภาษาการคำนวณทางลอจิก ซึ่งจะมีปริมาณมากขึ้นไปตามจำนวนประเภทของประโยคที่จะแปลในขณะนี้เราสามารถสร้างประโยคภาษาไทยจากภาษากลางที่มีโครงสร้างต่างๆกัน ในระยะเริ่มต้นนี้ได้ประมาณ 200 ประโยค จากจำนวนคำศัพท์ในพจนานุกรมที่อยู่ในขั้นทดลองขณะนี้ประมาณ 500 คำ โดยได้รับความร่วมมือในการวิเคราะห์ความหมายของคำศัพท์ในรูปของภาษากลางจากงานสร้างพจนานุกรมฯ ของสถาบันเทคโนโลยีพระจอมเกล้าธนบุรี โดยทางงานนี้ได้แก้ไขปรับปรุงข้อมูลให้เหมาะสมกับการสร้างประโยคภาษาไทย และสำหรับตัวอย่างภาษากลางก็ได้มีการร่วมกัน สมมติขึ้นประมาณ 200 ประโยค ดังกล่าว โดยทางงานได้สร้างรูปแบบทางซอฟท์แวร์ เพื่อกำหนดภาษากลางในการอินพุตเข้าสู่ระบบการสร้างภาษาไทยของงานนี้ แต่อย่างไรก็ตาม โครงสร้างประโยคหรือประเภทของประโยคที่ทำการทดลองในขณะนี้ ยังมีรูปแบบที่ไม่ยุ่งยากนัก การวิจัยและพัฒนาในลำดับต่อไปจะมีการแก้ไขปรับปรุงระบบให้ดียิ่งขึ้น

## เอกสารอ้างอิง


[1] A.B. Tucker, "Current Strategies in Machine Translation Research and Development", in Machine Translation Theoretical and Methodological Issues, Cambridge University Press, London, 1987, PP. 22-44.

[2] C.J. Fillmore; "A Case for Case", Universals in Linguistic Theory, ed. Emmon Bach and R.T. Harms, New York 1968.

[3] J.Slocum,"A Survey of Machine Translation: Its History, Current Status, and Future Prospects", American Journal of Computational Linguistics, Vol.11, No.1 January - March 1985, PP. 1 - 17.

[4] M.Nagao, J.Tsujii, J.Nakamura; " The Japanese Government Project for Machine Translation",American Journal of Computational Linguistics, Vol.11, No.2 - 3, April - September 1985, PP.91 - 110.

[5] P. Thajchayapong et al., " A Thai-English Machine Translation", in Proceeding of Regional Symposium on Computer Science and Its Applications, At KMITL, Bangkok, Thailand 27-29 January 1987.

[6] S. Charmonman et al., " Microcomputer-based Thai-English Machine Translation for Thai Rice-Mill Correspondence", same with [3].

[7] กำชัย ทองหล่อ; "หลักภาษาไทย", สำนักพิมพ์บำรุงสาส์น พ.ศ.2530

[8] รายงานความก้าวหน้าที่ ศอ.บ-003 เรื่อง

1) รายงานผลการวิจัย เลขที่ สจล. 01/2531 เรื่อง " Case Relation of English Language in VENUS Machine Translation System (Version 1.0) "

2) รายงานผลการวิจัย เลขที่ สจล. 02/2531 เรื่อง " การสร้างภาษาเป้าหมายในระบบการแปลภาษา VENUS "

3) รายงานผลการวิจัย เลขที่ สจล. 01/2532 เรื่อง " การพัฒนาและสร้างโปรแกรมเพื่อการแปลงคำสั่งของระบบ VENUS ในส่วนของการสร้างภาษาเป้าหมาย (ระยะที่ 1) "

# โครงการพัฒนาพจนานุกรมเพื่อการแปลภาษาด้วยคอมพิวเตอร์

ผศ.นวลทิพย์ ตันติเศวตรัตน์
นาง กุลวดี แหยมเกตุ
นางสาว ธีรนุช โชคสุวณิช
นางสาว แก้วใจ จันทร์เจริญ
นางสาว มณฑิกา บริบูรณ์
นางสาว นิศานาถ แท่นนิล
นักวิจัยในโครงการพัฒนาพจนานุกรมเพื่อการแปลภาษาด้วยคอมพิวเตอร์
ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์ สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี

**1. บทนำ** พจนานุกรมอิเล็กทรอนิกส์ (Electronic Dictionary) คือ แหล่งข้อมูล (knowledge database) เพื่อให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติ ซึ่งอาจกล่าวได้ว่า พจนานุกรมอิเล็กทรอนิกส์หรือ ED เป็นพจนานุกรมสำหรับคอมพิวเตอร์ EDสำหรับการแปลภาษาด้วยคอมพิวเตอร์ (Machine Translation) ประกอบด้วยข้อมูลต่างๆของหน่วยคำ (Morphological Information) ข้อมูลทางด้านหน้าที่ทางไวยากรณ์ของคำ (Syntactic Information) และรายละเอียดทางด้านความหมาย (Semantic Information) ข้อมูลเหล่านี้อยู่ในรูปของโค้ต (code) และสัญลักษณ์ (symbol) ต่างๆ การอธิบายความหมายของคำใน ED จะต้องให้ความหมายที่แน่นอนชัดเจนเพื่อลดความกำกวม เนื่องจากคอมพิวเตอร์ มีความสามารถในการคิดจำกัด ไม่สามารถเข้าใจความหมายของคำที่ซับซ้อนหรือมีความเหลื่อมกันแล้วแต่กรณีการใช้ดังเช่นมนุษย์ได้

**2. ชนิดของพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย** การพัฒนาพจนานุกรมอิเล็กทรอนิกส์เพื่อโครงการแปลภาษาแบบหลากภาษา (Multilingua Machine Translation) ของ NECTEC และ CICC เป็นการพัฒนา ED ภาษาไทย แบ่งออกเป็นชนิดต่างๆ ดังนี้ :

**2.1 พจนานุกรมศัพท์ทั่วไป** (Basic Term Dictionary) เป็นพจนานุกรมภาษาไทยสำหรับศัพท์ที่ใช้อยู่ในชีวิตประจำวัน ตั้งแต่ปี พ.ศ.2531 จนถึงปัจจุบันโครงการพัฒนาพจนานุกรมได้พัฒนาศัพท์ทั่วไปเสร็จสิ้นแล้วตามเป้าหมายที่กำหนดไว้ คือ 50,000 คำ

การคัดเลือกคำศัพท์ทั้งหมดได้มาจากพจนานุกรมไทย - อังกฤษ หรือพจนานุกรมไทยทั่วไป และคลังข้อความที่โครงการได้สะสมไว้จำนวนประมาณ 35 Mbyte ซึ่งข้อความเหล่านี้ได้จาก บทความวิชาการสาขาต่างๆ โดยเฉพาะสาขาวิชาคอมพิวเตอร์ บทความทั่วไปจากหนังสือพิมพ์ และนิตยสารที่ได้พิจารณาว่า มีการใช้ภาษาไทยที่ได้มาตรฐานเป็นที่ยอมรับ เรื่องสั้น และนวนิยายต่างๆ คำศัพท์ทั้งหมดจะมีข้อมูลหลายชนิดเกี่ยวกับคำในรูปของโค้ตป้อนไว้ในเครื่องคอมพิวเตอร์แบบพีซีโดยใช้โปรแกรม dbase III plus และเก็บข้อมูลในรูปของ dbase file และ text file จากนั้น ได้นำ text file ที่ได้เปลี่ยนรูปแบบของข้อมูล (Data Format)ให้อยู่ในรูปแบบที่นำเข้าพจนานุกรมของ Workstation โดยใช้รหัสสมอ ในรูปแบบที่บังคับเก็บไว้ใน Database File ของพจนานุกรมโดยใช้โปรแกรมเก็บข้อมูลเป็นภาษา C ข้อมูลในพจนานุกรมนี้สามารถนำมาแก้ไขได้โดยใช้โปรแกรม Dictionary Editor และข้อมูลในพจนานุกรมนี้พร้อมให้โครงการวิเคราะห์ภาษาไทยและโครงการสร้างภาษาไทยนำไปใช้ในงานแปลภาษาด้วยคอมพิวเตอร์ได้โดยใช้ซอฟท์แวร์ของระบบการแปลภาษานี้

**2.2 พจนานุกรมศัพท์เทคนิค** (Technical Term Dictionary) เป็นการพัฒนาศัพท์เทคนิคภาษาไทย ซึ่งเน้นทางด้านศัพท์คอมพิวเตอร์และอิเล็กทรอนิกส์ โดยมีจำนวนคำศัพท์ตามเป้าหมาย่ 25,000 คำศัพท์ คำศัพท์ทั้งหมดรวบรวมจากพจนานุกรมศัพท์เทคนิคภาษาอังกฤษและไทยทางด้านคอมพิวเตอร์จำนวนมาก และคลังข้อความจากบทความ และตำราทางด้านคอมพิวเตอร์ และสาขาวิชาต่างๆ ที่เกี่ยวกับคอมพิวเตอร์ ข้อมูลของคำใช้ระบบวิเคราะห์เช่นเดียวกับการพัฒนาพจนานุกรมศัพท์ทั่วไป การเก็บข้อมูลอยู่ในรูปแบบเช่นเดียวกับพจนานุกรมศัพท์ทั่วไปด้วย

**2.3 พจนานุกรมคำเกิดร่วม** (Cooccurrence Dictionary) เป็นพจนานุกรมให้ข้อมูลเกี่ยวกับการเกิดร่วมของคำในประโยคหรือในวลี เช่น เขียวขจี ขว้างจักร ข้อมูลนี้ช่วยให้โครงการวิเคราะห์ภาษาไทยเลือกมโนทัศน์ของคำ (concept) แทนให้กับคำศัพท์ภาษาไทย เมื่อคำศัพท์ 1 คำ มีหลายความหมาย หรือให้โครงการสร้างภาษาไทยสามารถเลือกคำที่ถูกต้องนำมาใช้ในการแทนรูปของมโนทัศน์ได้เมื่อสร้างภาษาไทยจากภาษากลาง (Interlingua) ในกรณีที่หนึ่งมโนทัศน์ (Concept) มีหลายรูปคำศัพท์ การพัฒนาพจนานุกรมคำเกิดร่วมนี้เพิ่งเริ่มดำเนินงานในต้นปี 2536 นี้

โดยเริ่มเก็บข้อมูลของการเกิดร่วมของคำต่างๆ และจะได้วิเคราะห์ และตั้งเป็นกฎโครงสร้างของคำเกิดร่วมต่อไป

**3. เนื้อหาของพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย** ในระบบการแปลภาษาด้วยคอมพิวเตอร์ระหว่าง NECTEC - CICC นี้ โครงการพัฒนาพจนานุกรมอิเล็กทรอนิกส์ได้ดำเนินการพัฒนาข้อมูลของคำเพื่อให้เหมาะสมกับทฤษฎีภาษาศาสตร์ที่ใช้ในโครงการแปลภาษานี้ คือ ทฤษฎี Phrase Structure Grammar, Case Grammar และ Dependency Grammar ตั้งแต่เริ่มโครงการในปี 2532 จนถึงปัจจุบัน มีการปรับปรุงข้อมูลเกี่ยวกับคำศัพท์ 2 ครั้ง เพื่อให้เหมาะสมกับการประมวลผลในการวิเคราะห์ และการสร้างภาษายิ่งขึ้น ข้อมูลที่รับการปรับปรุงทุกชนิดได้รับจากการปรึกษาหารือ และการเสนอแนะจากโครงการวิเคราะห์ภาษาไทย และโครงการสร้างภาษาทั้งสิ้น ข้อมูลหลักที่ให้กับคำศัพท์ต่างๆ มีดังนี้

**3.1 ข้อมูลเกี่ยวกับหน่วยคำ** (Morphological Information) เป็นข้อมูลเกี่ยวกับชนิดของคำ เช่น คำเดี่ยว หรือคำประสม ตัวสะกดของคำ การสืบเนื่องของคำ คำหลัก และตำแหน่งของคำหลัก สำหรับคำประสมข้อมูลส่วนนี้ช่วยการแบ่งคำในประโยค เพื่อช่วยในการหาคำในพจนานุกรม

**3.2 ข้อมูลทางวากยสัมพันธ์** (Syntactic Information) เป็นข้อมูลเกี่ยวกับชนิดหลักและชนิดย่อยของคำ (CATEGORY และ SUBCATEGORY) แบบโครงสร้างทางวากยสัมพันธ์ของคำกริยา (VERB PATTERN) รายละเอียดของข้อมูลหลักมีดังนี้

**3.2.1 ชนิดหลัก และชนิดย่อยของคำ** (Category และ Subcategory) มีดังนี้

CATEGORY : CAT คือ ชนิดของคำที่วิเคราะห์ตามหน้าที่ของคำในประโยค ในการแบ่งวิเคราะห์ชนิด และหน้าที่ของคำนี้ มุ่งให้คอมเพิวเตอร์ข้าใจภาษาเพื่อการประมวลผลโดยที่ตัวคอมพิวเตอร์มีความสามารถในการคิดจำกัด จึงมิได้แบ่งชนิดและหน้าที่ของคำตามหนังสือหลักไวยากรณ์โดยทั่วๆไป แต่นำมาจัดชนิด และหน้าที่ของคำใหม่ เพื่อให้ใช้ในการประมวลผลโดยคอมพิวเตอร์ได้ เช่น จัดให้คำซึ่งเดิมเป็น Adjective ให้เป็น Verb ชนิดหนึ่ง คือ VATT

คำนาม (Noun : N) คือ คำที่ใช้เรียกคน สัตว์ สิ่งของ

สรรพนาม (Pronoun : P) คือ คำที่ใช้แทนคำนาม และอาจแทนด้วยคำบ่งชี้ที่มีความหมายรวมถึงนามที่ละไว้ด้วย

กริยา (Verb : V) คือ คำที่แสดงกริยาอาการหรือสภาพของประธานในประโยค และถือเป็นคำหลักของประโยค

คำช่วยกริยา (Auxiliary Verb : AUX) คือ คำที่เกิดร่วมกับคำกริยาเพื่อแสดงกาล มาลา วาจก ลักษณาการ และเจตนาของประโยค และสามารถเกิดในตำแหน่งใดก็ได้

คำบ่งชี้ (Determiner : DET) คือ คำที่ใช้แสดงความเฉพาะเจาะจงของคำนามโดยปรากฏร่วมกับคำนามและคำลักษณะนาม

กริยาวิเศษณ์ (Adverb : ADV) คือ คำขยายกริยาเพื่อบอกว่าเหตุการณ์หรือสิ่งที่เกิดขึ้นนั้น เกิดเมื่อไหร่ ที่ไหน หรืออย่างไร ในภาษาไทยกริยาวิเศษณ์จะไม่ปรากฏหน้าคำกริยาที่ถูกขยายเลย

ลักษณะนาม (Classifier : CLAS) คือ คำที่ใช้บอกลักษณะเฉพาะของคำ เช่น คัน ฉบับ แท่ง เครื่อง อาจซ้ำกับคำๆนั้นเองก็ได้ เช่น ประเทศ อำเภอ ภาษา หรือบอกลักษณะที่เป็นกลุ่มของคำ เช่น ชุด คู่ คณะ กลุ่ม และรวมถึงคำที่แสดงหน่วย นับด้วย เช่น กิโลกรัม กระสอบ เมตร

คำสันธาน (Conjunction : CONJ) คือ คำที่ใช้เชื่อมวลี หรือประโยคเข้าด้วยกัน เพื่อแสดงความสัมพันธ์ในแง่ใดแง่หนึ่ง

คำบุรพบท (Preposition : PREP) คือ คำที่นำหน้าคำนามหรือคำสรรพนามเพื่อบอกตำแหน่งสถานที่ และความสัมพันธ์ระหว่างคำกริยาและคำนามหรือสรรพนามที่ตามมา

คำอุทาน (Interjection : INT) คือ คำที่ใช้ในการอุทาน

คำอุปสรรค (Prefix : FIXP) คือ คำที่ใช้ประกอบหน้าคำ เช่น กริยาเพื่อให้เป็นคำนามหรือคำกริยาวิเศษณ์ คำที่เติมคำอุปสรรคจะเปลี่ยนหน้าที่ของคำ โดยมิได้เปลี่ยนความหมาย เช่น การความ อย่าง โดย ตัวอย่าง การทำงาน ความรัก อย่างวิเศษ โดยตรง

คำลงท้าย (Ending : END) คือ คำที่ปรากฏท้ายประโยคเพื่อแสดงคำถามหรือแสดงความสุภาพ

คำปฏิเสธ (Negation : NEG) คือ คำที่แสดงการปฏิเสธหรือเพื่อให้ความหมายกลับเป็นตรงข้าม เช่น ไม่ ไม่ได้ มิได้ มิ ไม่ใช่

SUBCATEGORY : SUBCAT คือ ชนิดย่อยของคำเป็นการแบ่งประเภทของคำให้ย่อยลงไปอีกเพื่อแสดงลักษณะเฉพาะหรือความหมายที่แคบลงของคำเพื่อให้คอมพิวเตอร์วิเคราะห์ความแตกต่างระหว่างคำต่างๆ ได้ในระดับหนึ่ง ดังนี้

NOUN : N แบ่งชนิดย่อยได้ 7 ประเภท คือ

-Proper Noun : PRP หมายถึง วิสามัญนาม ซึ่งเป็นคำนามชี้เฉพาะ เป็นชื่อเฉพาะของคน หรือสิ่งใดสิ่งหนึ่ง แสดงความสัมพันธ์ในแง่การกที่เป็นชื่อเฉพาะ (Proper Noun) และเรียกว่า CASE NAME และไม่ได้มีหน้าที่เป็นคำขยาย (Attribute) เช่น กรุงเทพฯ มานะ อนุสาวรีย์ชัยสมรภูมิ

หมายเหตุ คำ "โตโยต้า" จะจัดให้เป็น NPRP เพราะถือเป็นชื่อเฉพาะ (NAME) แต่คำ "รถยนต์โตโยต้า" จะจัดให้เป็น NCMN เพราะคำ "โตโยต้า" เป็นเพียงคำขยาย (Attribute) ของ "รถยนต์" ไม่ได้มีการกสัมพันธ์เป็น NAME และมิได้มีเพียงหนึ่งเท่านั้น

-Cardinal Number : NCNM หมายถึง คำนามที่แสดงจำนวนนับ เช่น หนึ่ง สอง ร้อย ครึ่ง สามารถทำหน้าที่เป็นประธานหรือกรรมของประโยคได้ เช่น สามมากกว่าสอง

-Ordinal Number : NONM หมายถึง คำนามที่บอกลำดับที่ก่อน-หลัง เช่น ที่หนึ่ง ที่เท่าไร ที่สุดท้าย

-Adverb Like Noun : NALN หมายถึง คำนามที่ทำหน้าที่ได้เช่นเดียวกับคำกริยาวิเศษณ์(adverb) มักจะเป็นคำนามที่ใช้แสดงเวลา เช่น วันอาทิตย์ วันนี้ ปีที่แล้ว และเป็นคำนามที่มักปรากฏอยู่หลังกริยา และตำแหน่งต้นหรือท้ายในประโยค เช่นเดียวกับคำกริยาวิเศษณ์ (adverb) เช่น

ร้านปิดทุกวันอาทิตย์

เมื่อปีที่แล้วมีประชุมใหญ่ที่กรุงเทพฯ

-a Noun which is Conceptually an Adjective : NAJN หมายถึง คำนามที่แสดงการขยายคำอื่นๆ ได้เช่นเดียวกับคำวิเศษณ์ (Adjective) ได้แก่คำ ส่วนใหญ่ ส่วนมาก เช่น

ส่วนใหญ่เรียนจบชั้นอุดมศึกษา

-Label : NLBL หมายถึง คำนามที่เป็นเครื่องหมายหรือสัญลักษณ์ และทำหน้าที่ได้เหมือนคำนามทั่วๆ ไป เช่น

1,2,+,-,...

-Common Noun : NCMN หมายถึง สามัญนามซึ่งเป็นคำนามโดยทั่วๆไปที่ใช้เรียก คน สัตว์ สิ่งของ เช่น หนังสือ อาหาร หอ รวมทั้งคำนามที่เป็นประเภทย่อยของคำนามนั้นซึ่งไม่ได้แสดงการกสัมพันธ์เป็น NAME แต่แสดงการขยาย attribute เท่านั้น เช่น ดอกไม้ป่า รถยนต์นั่ง รถเก๋ง และคำนามที่มาจากการเติมคำอุปสรรค (Prefix) เช่น การกระทำ ความรัก ความสุข

PRONOUN : PRON แบ่งชนิดย่อยได้ 6 ประเภท

-Personal Pronoun : PPRS หมายถึง บุรุษสรรพนามซึ่งเป็นคำสรรพนามที่แทนตัวผู้พูด ผู้ที่พูดด้วย หรือผู้ที่กล่าวถึง เช่น กระผม ข้าพเจ้า เขา พวกเรา

-Demonstrative Pronoun : PDMN หมายถึง สรรพนามบ่งชี้ที่ใช้แทนสิ่งที่ผู้พูดชี้เฉพาะเจาะจงถึง เช่น

นี่คือ สิ่งที่ทุกคนต้องการ

โน่นคือ บ้านของเขา

-Definite Pronoun : PDEF หมายถึง สรรพนามชี้เฉพาะที่ใช้แทนเพื่อบอกขอบเขตของสิ่งที่พูดถึง ได้แก่คำ เหล่านี้ ทั้งหมด ทั้งคู่ ทั้งสอง เช่น

เราต้องการใช้ทั้งหมด

-Indefinite Pronoun : PIND หมายถึง สรรพนามไม่ชี้เฉพาะที่ใช้แทนคนหรือสิ่งทั่วๆ ไป ไม่มีขอบเขตที่แน่นอน ได้แก่คำ บ้าง ทั้งหลาย ทั้งมวล เช่น

บ้างชอบเดินไปทำงาน

-Interrogative Pronoun : PNTR หมายถึง สรรพนามที่แสดงการถาม ใช้แทนสิ่งที่ผู้พูดต้องการทราบคำตอบ ได้แก่คำ ใคร อะไร ที่ไหน อย่างไร เมื่อไร ทำไม เช่น

เราต้องใช้อะไรในการส่งข้อมูลระหว่างคอมพิวเตอร์

-Relative Pronoun : PREL หมายถึง ประพันธสรรพนามที่ใช้แทนนามที่อยู่ข้างหน้า และทำหน้าที่เชื่อมประโยคเข้าด้วยกัน ได้แก่คำ ที่ ซึ่ง อัน ผู้ เช่น

RC-232C เป็นสายที่ใช้ในการส่งข้อมูล

VERB : V แยกเป็น 3 ประเภท คือ

-Active Verb : VACT หมายถึง คำกริยาแสดงอาการของประธานในประโยค เช่น วิ่ง ให้ ประกาศ มอง เรียน คิด

-Stative Verb : VSTA หมายถึง คำกริยาแสดงสภาพ ความรู้สึกของประธาน หรือแสดงผลจากการกระทำอย่างใดอย่างหนึ่ง เช่น ราคา หนัก ลำบาก ได้ยิน เรียนรู้ เสียใจ เป็น คือ จำ

-Attributive Verb : VATT หมายถึง คำกริยาที่บ่งบอกคุณสมบัติของคำนาม มักมาจากคำคุณศัพท์นั่นเอง เช่น สวย ใหญ่ ผอม เย็น

AUXILIARY VERB : AUX แยกเป็น 5 ประเภท คือ

-Before Verb and Before "ไม่" : XVBM หมายถึง คำช่วยกริยาที่วางอยู่หน้ากริยาและหน้าคำปฏิเสธ "ไม่" ในประโยคปฏิเสธ เช่น จะ กำลัง อาจ(จะ) เกือบ(จะ) คล้ายจะ

-Before Verb and After "ไม่" : XVAM หมายถึง คำช่วยกริยาที่วางอยู่หน้า Verb และวางหลังคำปฏิเสธ "ไม่" ในประโยคปฏิเสธ เช่น ค่อย(จะ) น่า(จะ) ได้มัว (แต่จะ)

-Before Verb ; Before or After "ไม่" : XVMM หมายถึง คำช่วยกริยาที่อาจวางหน้า Verb และวางหน้าหรือหลังคำปฏิเสธ"ไม่"ในประโยคปฏิเสธ แต่ความหมายแตกต่างกันไป เช่น ควร(จะ) เคย(จะ) ต้อง อยาก(จะ) ชอบ(จะ) กรุณา อย่า โปรด ห้าม

-Before Verb ; At the beginning of sentence : XVBB หมายถึงคำช่วยกริยาที่อยู่หน้าประโยค เช่น โปรด กรุณา ห้าม อย่า

-After Verb ; At the End of Verb Phrase : XVAE หมายถึง คำช่วยกริยาที่วางหลังกริยาและวางไว้ท้ายกริยาวลี เช่น อยู่ อยู่แล้ว แล้ว ไป มา ขึ้น ลง เข้า ออก เสีย ไว้ เอา ให้ ดู

DETERMINER : DET แบ่งได้เป็นชนิดย่อย ดังนี้

-Definite Determiner : DED หมายถึง คำบ่งชี้ที่ใช้ชี้เฉพาะคำนาม และปรากฏอยู่ในตำแหน่ง ดังนี้

After Noun : DDAN คือ คำบ่งชี้ซึ่งวางอยู่หลังคำนาม หรือคำขยายนามที่ชี้เฉพาะนั้น แต่จะไม่มีคำลักษณะนาม (Classifier) ปรากฏอยู่ระหว่างคำนามกับคำบ่งชี้ชนิดนี้เลย เช่น นี่ นั่น โน่น นู่น

ตัวอย่าง รถนี่ราคาแพงมาก

After Noun and Classifier : DDAC คือ คำบ่งชี้ซึ่งวางอยู่หลังคำนามหรือคำขยายนามที่ชี้เฉพาะนั้น และจะมีคำลักษณะนาม (Classifier)วางอยู่ระหว่างคำนามกับคำบ่งชี้ชนิดนี้ได้ เช่น นี้ นั้น โน้น นู้น

ตัวอย่าง รถคันนี้ราคาแพงมา

Between Noun and Classifier : DDBC คือ คำบ่งชี้ซึ่งวางอยู่ระหว่างคำนามและคำลักษณะนาม หรืออยู่หน้าคำบอกจำนวน เช่น ทั้ง ตั้ง สัก อีก เพียง

ตัวอย่าง เครื่องทั้งชุด

เราต้องการเครื่องอีก 3 เครื่อง

After Quantitative Noun : DDAQ คือ คำบ่งชี้ที่วางหลังคำแสดงจำนวน เช่น เท่านั้น พอดี ถ้วน

ตัวอย่าง เรามีวิศวกรอยู่เพียง 3 คนเท่านั้น

-Indefinite determiner : DEIN หมายถึง คำบ่งชี้ที่ใช้แสดงจำนวน และปรากฏในตำแหน่ง ดังนี้

After Classifier : DIAC คือ คำบอกจำนวนที่วางอยู่หลังคำนามหรือคำขยายนามและอาจมีคำลักษณะนามอยู่ระหว่างคำนามและคำบอกจำนวนชนิดนี้ได้ เช่น ไหน ใด อื่น อื่นๆ ต่างๆ

ตัวอย่าง ที่นี่มีเครื่องชนิดต่างๆ ให้เลือก

Before Quantitative Noun : DIBQ คือ คำบอกจำนวนที่วางอยู่ระหว่างคำนาม และลักษณะนาม หรือหน้าจำนวน(ตัวเลข) เช่น บาง ราว เกือบ ประมาณ เกือบๆ ราวๆ

ตัวอย่าง เครื่องจะใช้เวลาประมาณ 3 นาที

After Quantitative Noun : DIAQ คือ คำบอกจำนวนที่วางอยู่หลังจำนวน เช่น เศษ เศษๆ กว่า กว่าๆ

ตัวอย่าง เครื่องจะใช้เวลาประมาณ 3 นาทีเศษ

Cardinal Number : DCNM คือ คำแสดงจำนวนนับ เช่น หนึ่ง ร้อย

ตัวอย่าง หนังสือสิบเล่มนี้

Ordinal Number : DONM คือ คำแสดงลำดับที่ เช่น ที่หนึ่ง ที่สิบ

ตัวอย่าง รถคันที่สิบ

ADVERB : ADV มีชนิดย่อย ดังนี้

-Adverb with Normal Form : ADVN หมายถึง คำกริยาวิเศษณ์ทั่วๆไป วางอยู่หลังคำกริยาที่ถูกขยาย เช่น จำเจ เร็ว เก่ง สม่ำเสมอ พอ เกิน ตรง

ตัวอย่าง เขาเดินตรงเข้าไปในบ้าน

-Adverb with Iterative Form : ADVI หมายถึง คำกริยาวิเศษณ์ที่เป็นคำซ้ำ (มีรูปซ้ำตัวเอง) เช่น เร็วๆ เสมอๆ มัวๆ

ตัวอย่าง เขาชอบมาที่นี่เสมอๆ

-Adverb with Prefixed Form : ADVP หมายถึง คำกริยาวิเศษณ์ที่มีคำอุปสรรค (อย่าง โดย...) ประกอบ เช่น อย่างรวดเร็ว โดยเร็ว

ตัวอย่าง ฉันจะไปที่นั่นโดยเร็วที่สุด

-Sentential Adverb : ADVS หมายถึง คำกริยาวิเศษณ์ที่ทำหน้าที่ขยายทั้งประโยค เช่น ปกติ ธรรมดา ทั่วไป ทั่วๆไป โดยธรรมดา ตามปกติ

ตัวอย่าง โดยทั่วไปจะมีฝนตกชุกในภาคใต้ของประเทศไทย

CONJUNCTION : CONJ แบ่งเป็นชนิดย่อยได้ 3 ประเภท

-Coordinating Conjunction : CCRG หมายถึง คำเชื่อมที่ใช้เชื่อมประโยคความรวม (Compound Sentence) หรือเชื่อมประโยคความเดียว (Simple Sentence) 2 ประโยคเข้าด้วยกัน เพื่อแสดงความสัมพันธ์ในลักษณะใดลักษณะหนึ่ง เช่น และ(เชื่อมความคล้อยตามกัน) แต่(เชื่อมความขัดแย้งกัน) หรือ(เชื่อมความให้เลือกเอาอย่างใดอย่างหนึ่ง)

-Comparative Conjunction : CCMP หมายถึง คำเชื่อมที่ใช้เชื่อมประโยคที่แสดงความเปรียบเทียบ เช่น กว่า ที่สุด

-Subordinating Conjunction : CSBR หมายถึง คำเชื่อมที่ใช้เชื่อมประโยค

ความซ้อน (Complex Sentence) เป็นการเชื่อมประโยคกับ Clause เข้าด้วยกัน หรือเชื่อมประโยคหลักกับประโยคย่อยเข้าด้วยกัน เพื่อขยายส่วนใดส่วนหนึ่งของประโยคหลักให้ชัดเจนขึ้น

ตัวอย่าง เขาซื้อขนมที่เธอชอบ (ขยายกรรมให้ชัดเจน)

เด็กที่เธอเลี้ยงไว้เป็นเด็กดี (ขยายประธาน)

PREPOSITION :PREP แบ่งเป็น 2 ชนิด คือ

-RVBP : Verbal Preposition เป็นคำบุรพบทที่เป็นคำกริยาได้ด้วย

ตัวอย่าง เขาไปจากบ้าน (เป็นบุรพบท)

เขาจากเธอ (เป็นกริยา)

-RPRE : Preposition except "RVBP" หมายถึงคำบุรพบทอื่นๆ

ตัวอย่าง เสื้ออยู่ในตู้

ENDING : END แยกเป็น 2 ประเภท คือ

-EAFF : คือ คำลงท้ายในประโยคบอกเล่า เพื่อแสดงการตอบรับหรือแสดงความสุภาพ เช่น ครับ ค่ะ จ้ะ

-EITT : คือ คำลงท้ายประโยค เพื่อแสดงการถาม และทำให้ประโยคบอกเล่ากลายเป็นประโยคคำถามแบบตอบรับ ปฏิเสธ เช่น หรือ ไหม ใช่ไหม ไม่ใช่หรือ

**3.2.2 แบบโครงสร้างของประโยค** (Verb Pattern)

VERB PATTERN : VP เป็นการกำหนดแบบโครงสร้างของประโยคโดยใช้คำกริยาเป็นหลัก แบบโครงสร้างนี้จะแสดงว่าประโยคแต่ละแบบประกอบด้วยคำซึ่งมีหน้าที่ใดทางไวยากรณ์ และมีความสัมพันธ์ระหว่าง concept เช่นใดซึ่งรวมทั้งconcept ที่เป็นคำนามกับคำนาม กริยากับคำนาม และอื่นๆ และสำหรับแบบโครงสร้างของประโยคนั้น คำกริยาจะเป็นคำหลัก (head concept) และคำที่นำหน้าและตามหลังกริยา จะมีความสัมพันธ์อย่างไรกับคำกริยา โดยใช้การอธิบายความสัมพันธ์การก (case relation) เทียบไว้กับความสัมพันธ์ทางไวยากรณ์ แบบโครงสร้างของประโยคนี้ นักวิจัยในโครงการพัฒนาพจนานุกรมฯได้ร่วมมือกับนักวิจัยในโครงการวิเคราะห์ภาษาต้นฉบับฯ กำหนดแบบโครงสร้างขึ้นรวม 18 แบบ ดังนี้

ตารางที่ 2 : แบบโครงสร้างของประโยค

| VP NO. | VERB PATTERN | SAMPLES & TMAPS | SAMPLE FIELD |
|---|---|---|---|
| 1 | SUB V | เขาเดิน<br>SUB=OBJ | |
| 1 A | V DOB | มีดอกไม้บนโต๊ะ<br>DOB=OBJ | |
| 2 | SUB V AUX | ลูกหมาเกิดแล้ว<br>SUB=OBJ | FADV.{AUX, ADV} |
| 3 | SUB V ADP | เขาลุกขึ้นแล้ว<br>SUB=OBJ | ADP.{ขึ้น} |
| 4 | SUB V CMP2 | น้ำประกอบด้วยไฮโดรเจน และออกซิเจน<br>SUB=OBJ,CMP2=ELM | CMP2.{ด้วย} |
| 5 | SUB V DOB CMP3 | เขาซื้อขนมให้ลูก<br>SUB=AGT,DOB=OBJ,<br>CMP3=TAR | CMP3.{ให้} |
| 11 | SUB V DOB | เขากินข้าวเหนียว<br>SUB=AGT,DOB=OBJ | |
| 11 A | SUB V DOB ADP | เขาเก็บอาหารไว้ในตู้<br>SUB=AGT,DOB=OBJ | ADP.{ไว้} |
| 12 | SUB V DOB IOB | เขาแจกขนมเด็ก<br>SUB=AGT,DOB=OBJ,<br>IOB=TAR | |
| 13 | SUB V CMP2 | เขาโกรธที่เรามาสาย<br>SUB=EXP,CMP2=REA | CMP2.{ที่} |
| 16 | SUB V DOB COMP | เขาบอกเราว่าเขาง่วง<br>SUB=AGT,DOB=TAR,<br>COMP3=OBJ | CMP3.{ว่า} |

| VP NO. | VERB PATTERN | SAMPLES & TMAPS | SAMPLE FIELD |
|---|---|---|---|
| 18 | SUB V COMP2 CMP3 | พ่อทะเลาะกับแม่ที่ลูกสอบตก | CMP2.{กับ} |
| | | SUB=AGT,CMP2=PRT, | CMP3.{ที่} |
| | | CMP3=REA | |

**3.3 ข้อมูลทางด้านอรรถศาสตร์** (Semantic Information) ข้อมูลประเภทนี้คือ ระดับชั้นของความหมาย ข้อมูลการก (case) คำที่มีความหมายเหมือนกันและต่างกัน และความหมายของคำเทียบไว้กับภาษาอังกฤษ และเทียบไว้กับ concept ซึ่งอยู่ในรูปของกลุ่มตัวเลขแทน concept แต่ละ concept เรียกว่า ID Number ข้อมูลเกี่ยวกับการณ์ลักษณะ มุมมองของผู้พูด และอื่นๆ ซึ่งมีข้อมูลหลักดังรายละเอียดต่อไปนี้

**3.3.1 กลุ่มของคำตามระดับชั้น** (Word Hierarchy) เป็นการแบ่งคำตามระดับชั้นของความหมาย คำหลักแต่ละคำในประโยค เช่น คำนาม คำกริยา คำวิเศษณ์ จะได้รับการจัดแบ่งกลุ่มตามความหมาย ซึ่งเรียกว่า AKO หรือ A Kind Of

AKO (A Kind of) เป็นข้อมูลที่ใช้จัดความหมายของคำที่มีอยู่ในภาษาให้เป็นหมวดหมู่ เพื่อเป็นประโยชน์ในการวิเคราะห์ความหมายของประโยค หรือเป็นการช่วยลดความกำกวมที่อาจเกิดขึ้นได้ในประโยค ดังตัวอย่างเช่น "ตากลับบ้าน"ประโยคนี้จะเกิดความกำกวมขึ้นที่คำนาม "ตา" ซึ่งเครื่องคอมพิวเตอร์ไม่สามารถบอกได้ว่า คำว่า "ตา" จะมีความหมายว่า "grandfather" หรือ "eye" จึงต้องอาศัยคำที่อยู่ใกล้เคียงมาช่วยในการตัดสิน ซึ่งก็คือคำกริยา "กลับ" จะพบว่า ข้อมูลคำกริยาตัวนี้ ต้องการประธานที่เป็นสิ่งมีชีวิต ซึ่งในที่นี้ก็คือ "ตา" ที่มีความหมายว่า "grandfather" ดังนั้นจึงจำเป็นต้องมีการแบ่งกลุ่มทางความหมายให้กับคำแต่ละคำโดยจัดหมวดหมู่ของความหมายเป็นลำดับขั้น จากกลุ่มที่ใหญ่จนถึงกลุ่มย่อยซึ่งมีดังต่อไปนี้

1 Concrete : รูปธรรม แบ่งเป็น

11 Subject : บุคคลและกลุ่มที่ประกอบด้วยคน แบ่งเป็น

111 Person : บุคคล เช่น ผู้ปกครอง เขา ผู้ดูแล คู่รัก เพื่อน

112 Organization : กลุ่มหรือองค์กรต่างๆ เช่น บริษัท กรรมการ กลุ่ม คณะ

12 Concrete place : สถานที่ แบ่งเป็น

121 Region : สถานที่เป็นดินแดนหรือมีขอบเขตที่แน่นอน เช่น จังหวัด

อำเภอ หมู่บ้าน

122 Nature : สถานที่เป็นธรรมชาติ แบ่งเป็น

1221 Topography : สถานที่ที่เป็นภูมิประเทศ เช่น ทุ่ง เทือกเขา ชายทะเล

1222 Astronomy : สถานที่ที่เกี่ยวข้องกับดาราศาสตร์ เช่น โลก พระอาทิตย์

123 Man-made place : สถานที่ที่มนุษย์สร้างขึ้น เช่น ทางด่วน ตึก บ้าน

13 Concrete thing : สิ่งของ แบ่งเป็น

131 Nature thing : สิ่งของที่เกิดตามธรรมชาติ แบ่งเป็น

1311 Living thing : สิ่งมีชีวิต

13111 Animal : สัตว์ เช่น หนอน หมา แมว

13112 A part of animal or its derivative : ส่วนของสิ่งมีชีวิต หรือสิ่งที่ได้จากสิ่งมีชีวิต เช่น มือ เท้า แขน ไข่ น้ำตา

13113 Plant : พืช เช่น ดอกไม้ มะนาว กุหลาบ

13114 A part of plant or its derivative : ส่วนของพืชหรือสิ่งที่ได้จากพืช เช่น กิ่งไม้ ราก เปลือก

1312 Substance : สสาร แบ่งเป็น

13121 Solid : ของแข็ง เช่น ทอง เงิน เพชร

13122 Liquid : ของเหลว เช่น น้ำ ปรอท น้ำมัน

13123 Gas : ก๊าซ เช่น อากาศ ไอน้ำ ก๊าซธรรมชาติ

132 Man-made thing : สิ่งของที่มนุษย์สร้างขึ้น แบ่งเป็น

1321 Material : วัตถุดิบ เช่น พลาสติก ผ้า

1322 Finished product : ผลิตภัณฑ์ โต๊ะ ดวงไฟ รถยนต์

1323 A part of product or its derivative : ส่วนของผลิตภัณฑ์หรือผลที่ได้จากผลิตภัณฑ์ เช่น ขาโต๊ะ ขี้ยางลบ ใบมีด

2 Abstract : นามธรรม แบ่งออกเป็น

21 Absrtact matter : นามธรรมที่ครอบคลุมเหตุการณ์และกริยาอาการต่างๆ แบ่งเป็น

211 Activity : กริยาและอาการต่างๆ แบ่งออกเป็น

2111 Action : การกระทำ เช่น ปฏิกริยา เล่น เดินทาง ใช้

2112 Mental action : การกระทำที่เกี่ยวกับความรู้สึกนึกคิด แบ่งเป็น

21121 Rational action : การกระทำที่ต้องใช้เหตุผลและความคิด เช่น คิด ตัดสินใจ ควบคุม

21122 Emotional action : การกระทำที่เกี่ยวข้องกับความรู้สึก เช่น รัก หวัง โกรธ เสียใจ

212 Phenomenon : ปรากฏการณ์ แบ่งออกเป็น

2121 Event : เหตุการณ์ เช่น ดวง โชค ความสำเร็จ

2122 Natural phenomenon : ปรากฏการณ์ธรรมชาติ เช่น พายุฝน สุริยุปราคา

2123 Social phenomenon : ปรากฏการณ์ทางสังคม เช่น การประท้วง สงคราม การจราจล

2124 Physiological phenomenon : ปรากฏการณ์ทางสรีระหรือทางร่างกาย เช่น อาการป่วย การเกิด การตาย การย่อยอาหาร

22 Abstract thing : นามธรรมที่มีลักษณะคล้ายสิ่งที่สามารถจับต้องได้ แบ่งเป็น

221 Product of activity : ผลของกริยาและอาการต่างๆ แบ่งออกเป็น

2211 Result of action : ผลของการกระทำ เช่น คำอธิบาย แผนการณ์ ข้อเสนอ

2212 Result of mental action : ผลของการกระทำที่เกี่ยวข้องกับความรู้สึกนึกคิด เช่น ความจำ ความเป็นห่วง ความรัก

2213 Abstract thing fixed as socially : สิ่งที่เป็นนามธรรมที่เกี่ยวข้องกับสังคม เช่น งานปีใหม่ งานลอยกระทง

2214 Abstract thing fixed as culture : สิ่งที่เป็นนามธรรมที่เกี่ยวข้องกับวัฒนธรรม เช่น ภาษาไทย รำวง มวยไทย

2215 Immaterial man-made thing : สิ่งที่ไม่ใช่วัตถุที่มนุษย์สร้างขึ้น
เช่น วรรณคดี เพลง บทความ โปรแกรมคอมพิวเตอร์

23 Abstract relation : ความสัมพันธ์เชิงนามธรรม แบ่งออกเป็น

231 Place : สถานที่ เช่น ขวา ซ้าย บน แถวๆ

232 Time : เวลา เช่น เมื่อวานนี้ สัปดาห์นี้ สมัยธนบุรี

234 Set : กลุ่ม เช่น เป็นแถว เรียงราย เป็นกลุ่ม

235 Attribute : คุณสมบัติของวัตถุหรือสิ่งของ แบ่งเป็น

2351 Quality : คุณลักษณะ เช่น ลักษณะ อุปนิสัย

2352 Condition : สภาวะ เช่น สภาพร่างกาย สถานการณ์

2353 Appearance : ลักษณะที่ปรากฏให้เห็น เช่น รูปร่าง สี ลาย

2354 Sort/Type : ประเภทหรือชนิด เช่น แบบ ชนิด ประเภท เผ่า

2355 Number/Quantity : จำนวน เช่น ปริมาตร ความสูง

2356 Relation : ความสัมพันธ์ เช่น ความเกี่ยวข้อง

2357 Function : หน้าที่และประสิทธิภาพ เช่น ความสามารถ
ประสิทธิภาพ คุณภาพ

236 Value : ค่าของคุณสมบัติหรือสิ่งของ แบ่งเป็น

2361 Quality : ค่าที่แสดงลักษณะ เช่น ใจดี สูง ง่าย

2362 Condition : ค่าที่แสดงสภาวะ เช่น เป็นอันตราย ปลอดภัย

2363 Appearance : ค่าที่แสดงลักษณะที่ปรากฏ เช่น กลุ่ม สี่เหลี่ยม
สว่าง แดง

2364 Sort/Type : ค่าที่แสดงชนิดหรือประเภท เช่น รุ่น2503 แบบดิจิตัล

2365 Number/Quantity : ค่าที่แสดงจำนวนหรือปริมาตร เช่น 10 วัน
มาก นาน

2366 Relation : ค่าที่แสดงความสัมพันธ์ เช่น ใกล้เคียง คล้าย ติดกัน

2367 Function : ค่าที่แสดงหน้าที่และประสิทธิภาพ เช่น (ประสิทธิภาพ)
สูง (คุณภาพ)เยี่ยม

หมายเหตุ คำ 1 คำ อาจมีหลาย AKO ได้ เนื่องจากคำแต่ละคำอาจมีมากกว่า 1 concept และ AKO จะพิจารณาจากความหมายของคำก็คือ concept นั่นเอง

ถ้าคำนั้นมีหลาย concept ก็จะต้องมีหลาย AKO ด้วย เช่นคำว่า "โรงงาน" จะมี 2 AKO คือ 112 คือ organizationและ 123 คือ man-made place

**3.3.2 ข้อมูลเกี่ยวกับความการกสัมพันธ์** (Case Relations) เป็นการแสดงความสัมพันธ์ระหว่างโครงสร้างทางวากยสัมพันธ์ และโครงสร้างทางอรรถศาสตร์ ซึ่งแสดงโดยใช้การกสัมพันธ์ระหว่างคำนามกับคำกริยา และคำอื่นๆในประโยคแปลงรูปโครงสร้างระดับพื้นผิวทางด้านไวยากรณ์ไปเป็นโครงสร้างความหมายระดับลึก เพื่อช่วยในการวิเคราะห์ประโยคภาษาไทยซึ่งได้ทำการวิเคราะห์ประโยคจากรูปพื้นผิวเก็บไว้เป็นโครงสร้างของความหมายระดับลึกของประโยคในรูปโครงสร้างของภาษากลาง (Interlingua structure) และงานสร้างภาษาจะรับโครงสร้างภาษากลางนี้ทำการสังเคราะห์ย้อนกลับขึ้นเป็นภาษาเป้าหมายอีกทอดหนึ่งความสัมพันธ์ทางด้านการกนี้จะเทียบไว้กับความสัมพันธ์ทางด้านไวยากรณ์ในแบบโครงสร้างของประโยคดังปรากฏใน 3.2.2 ชุดของการกสัมพันธ์นี้เป็นชุดที่กำหนดขึ้นใช้สำหรับโครงการแปลภาษาของระบบ PIVOT ของบริษัท NEC ซึ่งใช้มาแล้วเป็นเวลามากกว่า 10 ปี และโครงการแปลภาษานี้ ได้นำมาใช้ตั้งแต่ต้น การกสัมพันธ์มีประมาณ 39 แบบ ดังนี้

ตารางที่ 3 : การกสัมพันธ์ของระบบ PIVOT

| NO. | CASE NAME | READING |
|---|---|---|
| 1 | OBJ | object |
| 2 | AGT | agent |
| 3 | EXP | experiencer |
| 4 | INS | instrument |
| 5 | REA | reason |
| 6 | MEA | means |
| 7 | LOC | location |
| 8 | TIM | time |
| 9 | SOR | source |
| 10 | TAR | target |
| 11 | PRT | participant |

| | | |
|---|---|---|
| 12 | MAT | material |
| 13 | ELM | element |
| 14 | POS | possessor |
| 15 | POF | part of |
| 16 | NAM | name |
| 17 | ATT | attribute |
| 18 | VAL | value of attribute |
| 19 | QUANT | quantity |
| 20 | NUM | number |
| 21 | BEN | beneficiary |
| 22 | GOL | goal |
| 23 | MOD | modifier |
| 24 | MODS | sentencial modifier |
| 25 | PAR | parallel |
| 26 | CON | connection |
| 27 | EQ | equality |
| 28 | TES | thesaurus |
| 29 | REF | reference |
| 30 | CAP | capacity |
| 31 | FCS | focus |
| 32 | CAS | case |
| 33 | CASA | case assumption |
| 34 | CMPQUANT | quantity of comparative |
| 35 | MODQUANT | quantity of modification |
| 36 | RELBASE | base of relativity |
| 37 | CAU | causer |
| 38 | NMOD | noun modifier |
| 39 | DUMCMP | dummy case for comparative |

**3.3.3 ข้อมูลเกี่ยวกับความหมายระดับมโนทัศน์ของคำ** (Word concept) ซึ่งเป็นส่วนสำคัญเพื่อนำไปเชื่อมกับภาษาอื่นๆ ในการแปลภาษาแบบหลากภาษาโดยคอมพิวเตอร์ จะทำการประมวลผลโดยสร้างคำพื้นผิวของภาษาเป้าหมายจากมโนทัศน์ (concept) นี้ มโนทัศน์แต่ละมโนทัศน์ อาจแสดงด้วยรูปพื้นผิว (sunface form) ได้หลายรูป หรือรูปพื้นผิวหนึ่งอาจมีมโนทัศน์ได้หลายมโนทัศน์ เช่น คำว่า "คน" อาจเป็นได้ทั้ง "stir" หรือ "human being" การแทนคำพื้นผิวในรูปของมโนทัศน์ (concept) นี้ จะใช้กลุ่มตัวเลขแต่ละกลุ่มเป็นตัวแทนความหมายหรือมโนทัศน์โดยในการทำงานของโครงการแปลภาษาด้วยคอมพิวเตอร์ของ NECTEC-CICC นี้ กำหนดร่วมกันระหว่างประเทศสมาชิกในโครงการ คือ ประเทศจีน อินโดนีเซีย มาเลเซีย ญี่ปุ่น และไทย ให้ใช้พจนานุกรมของ EDR (Electronic Dictionary Research Institute) เป็นพจนานุกรมกลาง โดยพจนานุกรม EDR นี้ประกอบด้วย Entry Word (English surface)คำอธิบายความหมายทั้งหมดของแต่ละคำ และกลุ่มของตัวเลขแทนความหมายแต่ละความหมาย เรียกว่า ID Number เช่น คำว่า "เครื่องกรอง" เมื่อเทียบกับความหมายจำนวนถึง 18 ความหมาย ภายใต้คำว่า "FILTER" แล้วจะเทียบได้กับ " an apparatus through which liquids can be passed so as to make them clean" และมี ID No. เป็น "3be7d8" การเทียบคำภาษาไทยกับ ID Number นั้น ได้กำหนดความคล้ายคลึงของความหมายไว้เป็น 5 ระดับ ดังนี้

ระดับที่ 1 คำไทยและคำอธิบายความหมาย มีความหมายเท่ากัน

ระดับที่ 2 คำไทยมีความหมายกว้างกว่า คำอธิบายความหมาย

ระดับที่ 3 คำอธิบายความหมายมีความหมายกว้างกว่าคำไทย

ระดับที่ 4 คำไทยและคำอธิบายความหมายมีความหมายหรือความสัมพันธ์ในระดับหนึ่ง

ระดับที่ 5 คำที่มีความหมายเฉพาะในภาษาไทย เนื่องจากเกี่ยวข้องทางวัฒนธรรมไทยเท่านั้น

**3.3.4 ข้อมูลที่แสดงความเห็นของผู้พูด** (Speaker's Attitude) เป็นข้อมูลที่ผู้พูดแสดงทัศนคติที่มีต่อเหตุการณ์ เช่น ผู้พูดเห็นว่าเหตุการณ์คงจะเกิดขึ้น หรือเหตุการณ์ควรจะเกิดขึ้น สำหรับภาษาไทยเจตนคติของผู้พูดนี้แสดงด้วยหน่วยคำ (lexicon) ที่ปรากฏในตำแหน่งหน้าคำกริยา ในพจนานุกรมอิเล็กทรอนิกส์เพื่อการแปลภาษาด้วยคอมพิวเตอร์

หน่วยคำที่มีความหมายในการแสดงเจตนคตินี้จะเป็น ENTRYที่จัดอยู่ในAUXILIARY VERB CATEGORY

ตัวอย่าง : เขาน่าจะมาถึงได้แล้ว

(ผู้พูดมีความเห็นว่าเหตุการณ์ควรจะเกิดขึ้นแล้ว)

ตารางที่ 4 : การแสดงความเห็นของผู้พูดที่มีต่อเหตุการณ์

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| XLOOK | ดูเหมือนว่า,คล้ายจะ | SEEM,LOOK LIKE-ING |
| XSEEM | ราวกับว่า,ปรากฏว่า | SEEM AS IF,APPEAR AS IF |
| XINFE | จะ | WILL,MAY,MIGHT |
| XPRED | ต้อง,จง | MUST |
| XGUES | คง,คงจะ,อาจ,อาจจะ | MAY,MIGHT,COULD |
| XNEED | จำเป็น..ต้อง | NEED,MUST,HAVE TO |
| XPROH | อย่า,ห้าม | MUST NOT,SHOULD NOT |
| XPERM | นะ | MAY,CAN,COULD |
| XRECO | น่า,น่าจะ | SHOULD,BETTER, HAD BETTER |
| XADVI | ควร,ควรจะ | SHOULD |
| XCONCLU | ได้ | |

**3.3.5 ข้อมูลเกี่ยวกับมุมมองของผู้พูดที่มีต่อเหตุการณ์** (Speaker's Perspective) มุมมองของผู้พูดที่มีต่อเหตุการณ์ เป็นการให้ค่าของผู้พูดในฐานะที่ผู้พูดเป็นผู้มองเหตุการณ์ เช่นเดียวกับ XMDL สำหรับในภาษาไทย มุมมองของผู้พูดแสดงได้ด้วยหน่วยคำ (lexicon) ที่เกิดร่วมกับคำกริยา ซึ่งเมื่ออยู่ในพจนานุกรมอิเล็กทรอนิกส์เพื่อการแปลภาษาด้วยคอมพิวเตอร์จะอยู่ในกลุ่ม AUXILIARY VERB CATEGORY

ตัวอย่าง : เขาไม่เต็มใจมา

(ผู้พูดมองเหตุการณ์ที่เกิดขึ้นว่าผู้กระทำเหตุการณ์ไม่เต็มใจกระทำ)

ตารางที่ 5 :การแสดงมุมมองของผู้พูดที่มีต่อเหตุการณ์(SPEAKER'S PERSPECTIVE)

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| XTEND | มัก | TEND TO,BE APT TO |
| XESTIM | ควรค่า | BE WORTHY OF,DESERVE |
| XPOSS | สามารถ,..ได้,สามารถ..ได้ | CAN,BE ABLE TO,COULD |
| | ...เป็น | |
| XINVI | เถอะ,กัน,กันเถอะ | LET'S |
| XPLAN | ตั้งใจ,วางแผน | INTEND,PLAN |
| XWILL | เต็มใจ | WILL |
| XWISH | ประสงค์,หวัง,ต้องการ | WANT,WISH,DESIRE,HOPE |
| | ปรารถนา | |
| XPASS | ถูก,โดน,ได้รับ | BE-ED |
| XPOLI | โปรด,เชิญ,กรุณา | |
| MANY^OF^FREQ | มัก | OFTEN |
| XDECIDE | ตัดสินใจ | DECIDE |
| XTRY | ลอง...ดู | TRY-ING |
| XUNWILLING | ไม่เต็มใจ | BE UNWILLING |
| XEAGER | อยาก | BE EAGER TO |
| XFORGET | ลืม | FORGET |
| XACCUSTOMED | ชิน | BE ACCUSTOMED TO-ING |
| XMISTAKE | ผิด | MISTAKE |
| XFEEL | รู้สึก | BE SENSE OF |
| XDARE | กล้า | DARE |
| XPEC | เคย | USED TO |
| XBUSY | มัว | NEED,MUST,HAVE TO |
| XNOT^AT^ALL | ไม่...เลย | MUST NOT,SHOULD NOT |
| | | MAY,CAN,COULD |
| XENC | ซิ | SHOULD,BETTER,HAD BETTER |

**3.3.6 ข้อมูลด้านการมองเวลา** ที่เหตุการณ์เกิดขึ้นจากตัวเหตุการณ์ เรียกว่าการณ์ลักษณะ (Aspects) แบ่งเป็นการอธิบายโดยละเอียด คือ

ASPECT1 : EASP1 เป็นการณ์ลักษณะที่แสดงจุดเวลาภายในเหตุการณ์ เช่น จุดเริ่มต้น จุดที่เหตุการณ์กำลังดำเนินอยู่ จุดสิ้นสุดของเหตุการณ์ เป็นต้น

ASPECT2 : EASP2 เป็นการณ์ลักษณะที่แสดงช่วงเวลาภายในเหตุการณ์ เช่น ช่วงที่เหตุการณ์เกิดขึ้นและกำลังดำเนินอยู่ ช่วงที่เหตุการณ์เกิดขึ้นแล้วและสิ้นสุดลงแล้ว เป็นต้น

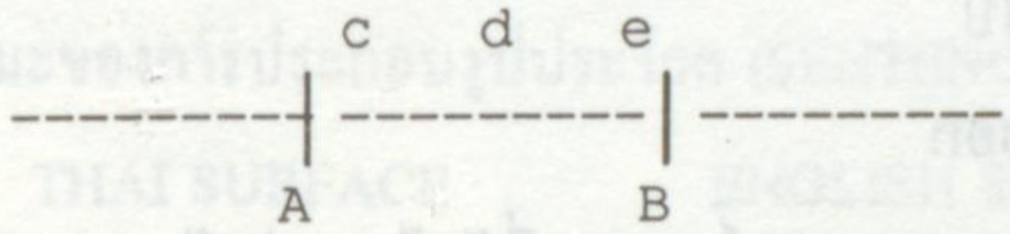
ตัวอย่าง : เหตุการณ์ "กินข้าว" ดำเนินไปตั้งแต่ A-B

ณ จุด A เขาเริ่มกินข้าว (EASP1)

B เขากินข้าวเสร็จ (EASP1)

จาก C-D เขากำลังกินข้าวอยู่(EASP2)

ภาพที่ 5 : แสดงจุดเวลา และช่วงเวลาภายในเหตุการณ์ (EASP1 และ EASP2)




EASP1 และ EASP2 แสดงด้วยหน่วยคำหลายกลุ่มได้แก่ AUXILIARY VERB, ADV และ ENDING

ตารางที่ 6 : การแสดงจุดเวลาภายในเหตุการณ์ (EASP1)

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| XBEGIN | เริ่ม,เริ่มต้น,ตั้งต้น | BEGIN,START |
| XABOUT | กำลังจะ,เกือบจะ | BE ABOUT TO |
| XCONT2 | ..ต่อ,กำลัง...อยู่ | CONTINUE |
| XFINISH | ..เสร็จ,แล้ว,จบ | FINISH |
| XKEEP | ไว้ | |
| XLOSE | เสีย | |
| XRECENT | เพิ่ง | JUST |

| | | |
|---|---|---|
| XRECENT | เพิ่ง | JUST |

ตารางที่ 7 : การแสดงช่วงเวลาภายในเหตุการณ์ (EASP2)

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| PROG | ยัง,กำลัง | BE-ING |
| PERF | ได้,ได้...แล้ว | HAVE-EN |
| STAT | อยู่ | |
| CONTI | ..ไป | |
| XIN | เข้า | |
| XDOWN | ลง | |
| XUP | ขึ้น | |
| XFORWORD | มา | |
| XAWAY | ไป | |
| XOUT | ออก | |

**3.3.7 ข้อมูลด้านการมองเหตุการณ์** จากจุดที่ตัวผู้พูดอยู่ ผู้พูดมองเหตุการณ์ว่าเป็นสิ่งที่อยู่ในอดีต ในปัจจุบัน หรืออนาคต (TIME) ข้อมูลนี้ไม่ใช่สิ่งที่เด่นชัดในภาษาไทย แต่ก็แสดงได้ด้วย AUXILIARYVERB CATEGORY เช่น เหตุการณ์ที่เกิดขึ้นในอดีตแสดงด้วยหน่วยคำ "แล้ว" เหตุการณ์ที่จะเกิดขึ้นในอนาคต แสดงด้วยหน่วย "จะ" เป็นต้น

ตัวอย่าง : เขาจะไปต่างประเทศ

(จากจุดที่ผู้พูดอยู่เหตุการณ์ยังไม่เกิดขึ้น)

ตารางที่ 8 : การแสดงการมองเหตุการณ์จากจุดที่ตัวผู้พูดอยู่ (TIME)

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| PAST | ได้,เสร็จ,ไว้,แล้ว | PAST TENSE |
| PRES | DEFAULT | PRESENT TENSE |
| FUTU | จะ,กำลังจะ | FUTURE TENSE |

**3.3.8 ข้อมูลด้านลักษณะของการประกอบรูปประโยค** ซึ่งปรากฏเป็นประโยคบอกเล่า ประโยคคำถาม ประโยคคำสั่ง (SENTENCE STYLE) แต่เจตนาของผู้พูดอาจไม่สอดคล้องกับลักษณะประโยคก็ได้ เช่น ผู้พูดใช้ประโยคคำถามในการบอกเล่า ดังตัวอย่าง "ใครจะรอให้เสียเวลา" (ความหมายของผู้พูดคือผู้พูดบอกให้ทราบว่าผู้พูดไม่รอ )

ตัวอย่าง : เขากลับถึงบ้านแล้ว ( ประโยคบอกเล่า )

เขากลับถึงบ้านกี่โมง ( ประโยคคำถาม )

กลับบ้านเร็วๆ ( ประโยคคำสั่ง )

ลักษณะของประโยคแสดงด้วยหน่วยคำ(LEXICON) หลายกลุ่มเช่นเดียวกัน กล่าวคือ ประโยคคำถามประกอบด้วยคำในกลุ่ม PRONOUN "ใคร" "อะไร" เป็นต้น และ ENDING "หรือ" เป็นต้น ส่วนประโยคคำสั่งอาจประกอบด้วย AUXILIARY บางตัว เช่น โปรด กรุณา อย่า

ตารางที่ 9 : ลักษณะของการประกอบรูปประโยค (SENTENCE STYLE)

| CP | THAI SURFACE | ENGLISH SURFACE |
|---|---|---|
| DEC | ประโยคบอกเล่า | DECLARATIVE SENTENCE |
| IMP | ประโยคคำสั่ง<br>โปรด,เชิญ,กรุณา,จง<br>อย่า,ห้าม | IMPERATIVE SENTENCE |
| INT | ประโยคคำถาม<br>หรือ,ทำไม,อะไร,อย่างไร<br>ไหม,ที่ไหน,หรือไม่,เมื่อไร | INTERROGATIVE SENTENCE |

**3.3.9 ข้อมูลเกี่ยวกับการกำหนดการก** (Case) หรือค่าให้กับคำบุรพบทที่เป็น Field Entry ซึ่งมักจะเป็น Free Case คือ ความสัมพันธ์ที่ไม่ต้องเกิดในทุกกรณี หรือทุกประโยค เช่น บน มี SUBFEA คือ LOC = OON คำบุรพบท "บน" ถ้าเกิดในประโยคใด ก็จะมีการกสัมพันธ์ คือ LOC บอกสถานที่ และมีค่าเป็น OON

สรุปข้อมูลที่กำหนดให้แก่คำไทย มีดังนี้

ENTRY คือ คำในภาษาไทยที่กำหนดจากคำที่ใช้โดยทั่วไป (Basic Term) ในภาษาไทย หรือเป็นคำที่ใช้เฉพาะในแต่ละสาขาเป็นศัพท์เทคนิค (Technical Term) เพื่อนำมาวิเคราะห์ หรือสร้างคำในการแปลด้วยคอมพิวเตอร์ ตัวอย่างเช่น นักเรียน ผม คุณ รัก คอมพิวเตอร์ บัฟเฟอร์ เป็นต้น

TYPE คือ ลักษณะของคำ แบ่งเป็นคำเดี่ยว (Single Word) ใช้สัญลักษณ์ (S) แทนหมายถึง คำที่มีพยางค์เดียว เช่น พ่อ แม่ น้อง หรือคำหลายพยางค์ที่มีความหมายเดียว เช่น กุหลาบ กำมะถัน มะลิ และคำประสม หรือคำรวม (Compound Word) ซึ่งเกิดจากการนำคำต่างๆ มาประสมกัน มักจะประกอบด้วยคำหลักและคำขยาย หรือคำแก่นและคำเสริม หมายถึงคำที่เกิดจากการรวมความหมายเข้าด้วยกันมีความหมายหลักอยู่ที่คำใดคำหนึ่ง และใช้สัญลักษณ์ (C) ตามด้วยตัวเลขกำกับเพื่อบอกคำหลักและการตัดคำ เช่น ตู้กับข้าว (คำหลักคือ ตู้ คำขยายคือ กับข้าว เพื่อบอกชนิดของตู้) ใช้สัญลักษณ์ C11 เลข 1 ตัวแรก หมายถึง คำหลัก เลข 1 ตัวหลัง หมายถึง 1 อักขระแล้วแยกคำ ห้องสมุดดนตรี (คำหลักคือ ห้องสมุด คำขยายคือ ดนตรี เพื่อบอกลักษณะของห้องสมุด) ใช้สัญลักษณ์ C16

CAT : Category คือชนิดหลักของคำ เช่น คำนาม (N) คำกริยา(V) โค้ตค่าของ CATด้วยตัวย่อที่ปรากฏในคำอธิบายรายละเอียดของข้อมูลเกี่ยวกับ CATEGORY

SUBCAT : Subcategory คือชนิดย่อยของคำ เช่น คำนาม(NCMN) คำกริยาแสดงอาการ(VACT) สำหรับ CAT และ SUBCAT นี้ นักวิจัยในโครงการได้กำหนดขึ้นเพื่อให้เหมาะกับภาษาไทย เช่น รูปคำของคุณศัพท์(adjective)ในภาษาอังกฤษจะได้รับการกำหนดเป็นคำกริยาในภาษาไทย โดยมีประเภทย่อยของคำต่างจากกริยาอื่นๆ

CP : ID Number คือ ความหมายของคำ ซึ่งจะเป็นส่วนสำคัญที่นำไปเชื่อมต่อกับภาษาอื่น แล้วจึงดึงคำในภาษานั้นๆ ที่มีความหมายตรงกันออกมาใช้ในการแปลภาษาด้วยเครื่องคอมพิวเตอร์ ดังนั้น รูปคำในแต่ละภาษาจะเป็นอะไรก็แล้วแต่ภาษานั้นๆ แต่จะมีความหมายเหมือนกัน ในแต่ละภาษาอาจมีรูปคำหลายรูปที่แสดง CP เดียว และในทางตรงข้าม รูปคำเดียวถ้านำมาใช้หลายความหมายก็ อาจจะมี CP หลายตัวก็ได้ เช่น "เมตตา" และ "กรุณา" มี CP คือ be kind คำ "กบ" คำเดียวมี CP เป็น carpenter's plane, fill to overflowing หรือ frog ก็ได้ และ CP เหล่านี้จะถูกแทน

ด้วยกลุ่มตัวเลขอีกทอดหนึ่ง

SIM : ระดับความเหมือนของความหมายระหว่างคำไทย และ JD Number

ENG : English คือ คำเทียบเคียงที่เป็นภาษาอังกฤษ ใช้เพื่อช่วยในการหาความหมาย

AKO : A kind of คือ การให้ความหมายของคำเป็นหมวดหมู่ เช่น คน มี AKO เป็น 111 คำกริยาแสดงอาการมี AKO เป็น 2111 หมวดหมู่ของความหมายเป็นการกำหนดลำดับชั้นของคำตามความหมายรวมใหญ่ เช่น สิ่งมีชีวิต และไม่มีชีวิต จากสิ่งมีชีวิตจะแยกเป็นคน สัตว์ เป็นต้นซึ่งหมวดหมู่นี้จะต้องใช้ร่วมกันกับประเทศสมาชิกอื่นๆ ที่เข้าร่วมวิจัยในโครงการการแปลภาษาด้วยคอมพิวเตอร์กับ CICC ด้วย

SYN : Synonym คือ คำที่มีความหมายเหมือน คือคล้ายคลึงกับคำ entry เพื่อช่วยในการวิเคราะห์คำ เช่น คำ "กรุณา" มี SYN เป็น "เมตตา"

ANT : Antonym คือ คำที่มีความหมายตรงข้ามกับ ENTRY เพื่อช่วยในการเปรียบเทียบคำ เช่น คำ "ดี" มี ANT เป็น "เลว"

VPPAT : Verb Pattern คือ โครงสร้างของคำกริยาในระดับโครงสร้างผิว (SurfaceStructure) เพื่อแสดงตำแหน่งการปรากฏของคำกริยากับการเกิดร่วมกับคำอื่นๆ ในประโยค

VP : Verb Pattern Number คือ ลำดับโครงสร้างของคำกริยา เพื่อช่วยให้คอมพิวเตอร์เรียกค้นคำกริยาที่ต้องการวิเคราะห์ และสร้างได้รวดเร็วยิ่งขึ้น

MAPS : Mapping Sentence คือ การแสดงความสัมพันธ์ในแง่การกระหว่างคำนามกับคำกริยา และคำอื่นๆ ในประโยค TMAPS จะช่วยเปลี่ยนระดับจากระดับพื้นผิวของภาษา หรือระดับประโยค เป็นระดับความหมายในส่วนของการวิเคราะห์ และในทางตรงข้ามจะช่วยเปลี่ยนระดับจากความหมายไปเป็นระดับประโยคในส่วนของการสร้างภาษาเป้าหมายโดยใช้การกสัมพันธ์เป็นส่วนเชื่อมต่อที่ตรงกันของทั้ง 2 ระดับ เช่น คำกริยา "ซื้อ" มี TMAPS เป็น "SUB=AGT, DOB=OBJ"

CMP2, CMP3 คือ ตำแหน่งของคำบุรพบทที่ต้องเกิดกับคำกริยาในประโยค โดยปรากฏหลังคำกริยาตามลำดับในตำแหน่งที่ 2 และ 3 เช่น กริยา "ทะเลาะ" ต้องมีคำบุรพบท "กับ" เกิดขึ้นด้วยเสมอ เพื่อบอกผู้ร่วมทะเลาะ และปรากฏในตำแหน่งที่ 2

หลังกริยาจึงเป็น CMP2 "เขาทะเลาะกับน้อง" แต่กริยา "ซื้อ" มีคำบุรพบท "ให้" เกิดหลังกรรมตรง ในตำแหน่งที่ 3 จึงเป็น CMP3 "แม่ซื้อขนมให้ลูก"

ADP : Adverb Particle คือ คำกริยาวิเศษณ์ (คำขยายกริยา) ที่ประกอบหลังกริยาแล้วทำให้ความหมายของกริยานั้นเปลี่ยนไปจากเดิม เช่น "น้องลุกขึ้นแล้ว" ADP คือ "ขึ้น" ทำให้ ความหมายของกริยา "ลุก" เปลี่ยนเป็น "ตื่น"

SAMPLE คือ ตัวอย่างประโยคที่ใช้คำที่กำหนด เพื่อช่วยในการวิเคราะห์คำและการใช้คำนั้นๆในประโยค

XMDL : Modality คือ เจตนคติ(attitude)ของผู้พูด เช่น XNEED คือ "ต้อง" XGUES คือ "คง" "อาจ"

XEASP1 : Aspect1 คือลักษณาการในภาษาไทย โดย XEASP1 เป็นจุดของเวลา เช่น XBEGIN คือ "เริ่ม" "เริ่มต้น" XABOUT คือ "กำลังจะ" "เกือบจะ"

XEASP2 : Aspect2 คือลักษณาการในภาษาไทย โดย XEASP2 เป็นช่วงเวลา เช่น PROG คือ "ยัง" STAT คือ "อยู่"

XCHSC : Speaker's Perspective คือ สภาวะของประธาน (Subject) และผู้พูด(Speaker) ในประโยค เช่น XPOSS คือ สามารถ , ได้ XPASS คือ ถูก, โดน XPOLI คือ โปรด, กรุณา XINVI คือ เถอะ

XSSTL : Sentence Style คือเจตนาของประโยค เช่น DEC (Declarative Sentence ) คือประโยคบอกเล่า IMP (Imparative Sentence) คือประโยคคำสั่ง INT (Interrogative Sentence) คือประโยคคำถาม

TIM : TIME คือ กาลเวลาที่เกิดเหตุการณ์นั้น เช่น PRES (Present) ปัจจุบัน PAST (Past)อดีตกาล FUTU(Future) อนาคตกาล

SUBFEA : Sub feature คือ การกำหนด caseให้กับบุรพบทที่เป็นField Entry เช่น บน มี SUBFEA คือ LOC=0ON

TORGSPL คือ ตัวสะกดเป็นภาษาไทย มีค่าเท่ากับ ENTRY เพื่อช่วยกำหนดรูปคำในการสร้างภาษาเป้าหมายของ Generation

NUM : เป็นลักษณนามของคำ

NOM : การแปลงคำกริยาให้เป็นคำนาม โดยใส่ การ, ความ

**4. ปัญหาของการพัฒนาพจนานุกรมอิเล็กทรอนิกส์** งานพัฒนาพจนานุกรมของโครงการจนปัจจุบันนี้ อาจกล่าวได้ว่า ประสบผลสำเร็จไปเพียงขั้นแรกเท่านั้น จากการประเมินผลของการนำไปใช้ ในการแสดงการแปลภาษาด้วยคอมพิวเตอร์ระบบหลากภาษาซึ่งจัดขึ้น ณ กรุงโตเกียว ประเทศญี่ปุ่น เมื่อวันที่ 15-16 ธันวาคม 2535 นั้น ยังปรากฏว่าการเชื่อมคำด้วยมโนทัศน์ หรือ ID Number ยังไม่ประสบผลสำเร็จดังที่ควรจะเป็นเนื่องจากแต่ละภาษามีคำศัพท์ซึ่งบ่งว่าเป็นคำศัพท์เฉพาะของแต่ละภาษาเป็นจำนวนมาก ซึ่งทำให้การสร้างภาษาเป้าหมายของชาติต่างๆประสบปัญหาไม่สามารถจะแปลออกมาได้ ซึ่งปัญหานี้อาจสืบเนื่องจากพจนานุกรมกลาง คือ พจนานุกรมของ EDR ยังไม่สามรถจะคลอบคลุมความหมายของคำในภาษาต่างๆได้เพียงพอ หรือ ปัญหาอาจเกิดขึ้นจาก การเชื่อมคำด้วย ID Number ของนักวิจัยแต่ละชาติ ซึ่งยังไม่ได้ให้ความหมายทุกความหมายเท่าที่จะเป็นไปได้แก่มโนทัศน์ทุกมโนทัศน์ เพื่อเปิดโอกาสให้มีการเลือกคำพื้นผิวที่กว้างขึ้น

นอกจากการเชื่อมต่อของคำพื้นผิวและมโนทัศน์ดังกล่าวแล้ว ข้อมูลที่ให้แก่คำแต่ละคำยังต้องการการแก้ไขปรับปรุงอีกมาก เนื่องจากข้อมูลบางชนิดยังมีความกำกวมและยังไม่สมบูรณ์พอ ซึ่งเป็นผลจากการกำหนดกฎเกณฑ์ของข้อมูล ยังไม่คลอบคลุมการใช้ในภาษาไทย และเมื่อจำนวนคำเพิ่มมากขึ้น การแก้ไขข้อมูลยิ่งทวีความลำบากขึ้น

ลักษณะของภาษาไทยเองก็ทำให้เกิดปัญหาแก่การพัฒนาพจนานุกรมด้วย เนื่องจากคำไทยประสมมีลักษณะเป็นการเรียงคำเดี่ยวหลายคำเข้าด้วยกัน ดังนั้น การให้คำจำกัดความแก่คำไทย และการเก็บคำเข้าพจนานุกรมจึงเป็นงานที่ยากแก่การตัดสินใจ ซึ่งในปัจจุบันนี้โครงการใช้วิธีเก็บคำตามความเห็นชอบของนักวิจัยในโครงการเอง ซึ่งคำบางคำที่เก็บไว้จึงมีโครงสร้างเหมือนประโยค เช่น "เครื่องกรองน้ำ" ดังนั้นการนำพจนานุกรมไปใช้ในการตัดคำจึงเกิดปัญหาในบางครั้งและการเก็บคำในลักษณะนี้ ทำให้ขนาดของพจนานุกรมใหญ่เกินไปด้วย

**5. สรุป** พจนานุกรมอิเล็กทรอนิกส์ที่โครงการพัฒนาขึ้นนี้ ยังมีวัตถุประสงค์เฉพาะคือการนำไปใช้ในการประมวลผลภาษาเพื่อการแปลภาษาด้วยเครื่องคอมพิวเตอร์ระบบหลากภาษา การกำหนดคำและข้อมูลจึงเพื่อเอื้ออำนวยแก่การทำงานนี้โดยเฉพาะ การนำไปใช้นอกเหนือจากการแปลภาษาด้วยคอมพิวเตอร์ จึงต้องการการดัดแปลงให้

เหมาะสม ซึ่งโครงการเองก็มีจุดมุ่งหมายในการปรับปรุงข้อมูลให้ได้มาตรฐาน ในการใช้ทั่วไปอยู่แล้ว แต่เนื่องจากนักวิจัยของโครงการมีจำนวนจำกัด ดังนั้น การปรับปรุงนี้จึงต้องใช้เวลา การพัฒนาพจนานุกรมอิเล็กทรอนิกส์เพื่อการแปลภาษาด้วยคอมพิวเตอร์นี้ต้องทำงานร่วมกับกฎการวิเคราะห์และกฎการสังเคราะห์ภาษาไทยอย่างใกล้ชิด เนื่องจากพจนานุกรมเป็นฐานความรู้ในการประมวลภาษา ดังนั้น ข้อมูลที่ได้จากการนำไปใช้วิเคราะห์และการสร้างภาษาไทยจึงเป็นรากฐานของการปรับปรุงข้อมูลของโครงการพัฒนาพจนานุกรม ฯ นอกจากนั้นโครงการยังมีแผนการณ์พัฒนาข้อมูลอีกหลายอย่าง โดยอาศัยคลังข้อความขนาดใหญ่ที่โครงการรวบรวมไว้ในรูปของ Text file เช่น การเกิดของคำแสดงหน้าที่ (function words) ทั้งหลาย การแบ่งคำ การแบ่งความ และอื่นๆ ซึ่งจะทำให้เกิดการพัฒนาการกำหนดข้อมูลอีกมาก

# การสร้างระบบรวมเพื่อการแปลภาษาด้วยคอมพิวเตอร์

กัลยา นฤดมกุล
ภาควิชาคณิตศาสตร์ มหาวิทยาลัยมหิดล
บุญเจริญ ศิริเนาวกุล
อาณัติ เชิดชูชัย
วุฒิชัย พลวิเศษ
ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์
สถาบันเทคโนโลยีพระจอมกล้าธนบุรี

**บทคัดย่อ** การสร้างระบบรวมเพื่อการแปลภาษานี้เป็นการสร้างระบบคอมพิวเตอร์เพื่อให้ขบวนการของการแปลภาษาที่อยู่ต่างเครื่องกันสามารถทำงานร่วมกันได้ โดยอาศัยเครื่องคอมพิวเตอร์ตัวหนึ่งเป็นตัวรวมการทำงานของทั้งระบบ ตัวรวมนี้จะทำหน้าที่ส่งผ่านข้อมูลระหว่างเครื่อง พร้อมทั้งแสดงผลของการแปล ระบบที่สร้างขึ้นนี้ได้ทำการทดลองกับการแปลภาษาไทย อังกฤษ ญี่ปุ่น และ จีน โดยทำการแปลประโยคตัวอย่างของภาษาหนึ่งไปเป็นภาษาอื่นๆทุกภาษา ซึ่งให้ผล เป็นที่น่าพอใจ

**1. บทนำ** โครงการพัฒนาระบบรวมเพื่อการแปลภาษาด้วยคอมพิวเตอร์ (The Development of Integrated System for Machine Translation) เป็นหนึ่งในโครงการย่อยของโครงการหลัก ปัญญาประดิษฐ์ (Artificial Intelligence) โครงการนี้เริ่มตั้งแต่ปี พ.ศ. 2531 จนถึงปัจจุบัน

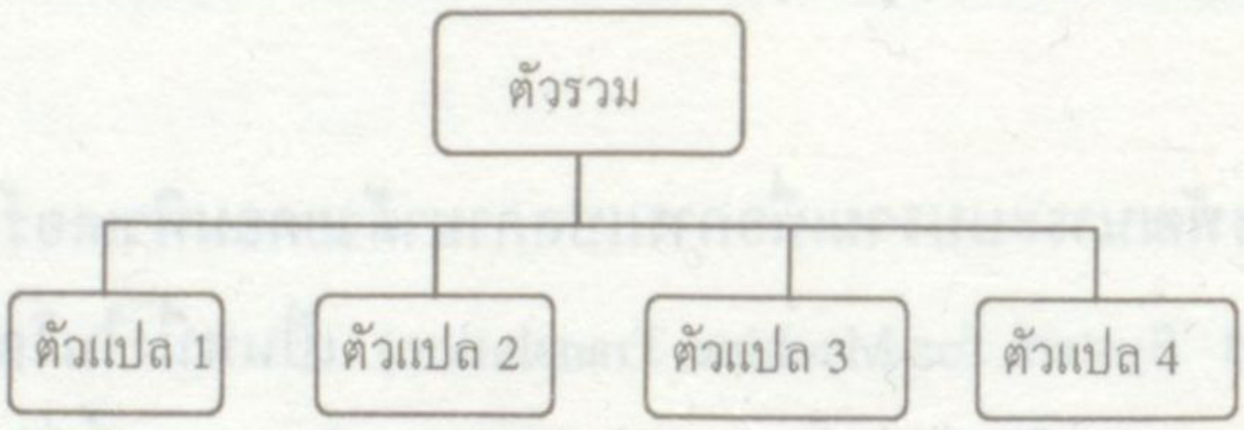
วัตถุประสงค์ของโครงการวิจัยนี้คือพัฒนาระบบรวมเพื่อการแปลภาษา ทั้งนี้เนื่องจากระบบการแปลภาษาด้วยคอมพิวเตอร์เป็นระบบที่มีขั้นตอนการประมวลผลค่อนข้าง ซับซ้อนและมีฐานความรู้ ฐานข้อมูลขนาดใหญ่อีกด้วย เช่น พจนานุกรม จำเป็นต้อง แบ่งขั้นตอนการประมวลผลออกเป็นมอดูล จากนั้นเชื่อมต่อการทำงานของแต่ละมอดูลเข้าด้วยกัน เรียกการเชื่อมต่อนั้นว่าเป็นการสร้างระบบรวม ในการพัฒนาระบบรวมนั้นจะต้องคำนึงถึง ความเร็วในการประมวลผล การใช้งานต้องง่ายและสะดวก การดูแลรักษาฐานข้อมูลต้องง่ายต่อการดูแลรักษาและปรับปรุง นอกจาก

นี้วิธีการเชื่อมต่อมอดูลจะต้องไม่ทำให้อินพุท-เอาท์พุทของแต่ละมอดูลเปลี่ยนแปลง นั่นคือต้องไม่มีผลกระทบต่อการประมวลผลในแต่ละมอดูล

โครงการวิจัยได้สร้างระบบรวม (Integrated System) ซึ่งสามารถเชื่อมต่อระบบการประมวลผลทั้งที่อยู่ในเครื่องเดียวกันและระหว่างเครื่องโดยเขียนโปรแกรมด้วยภาษาซีบนยูนิกซ์(Unix) และ X-Window การประมวลผลของระบบทำบนเครื่องสถานีงาน (Workstation) EWS 4800 [1]

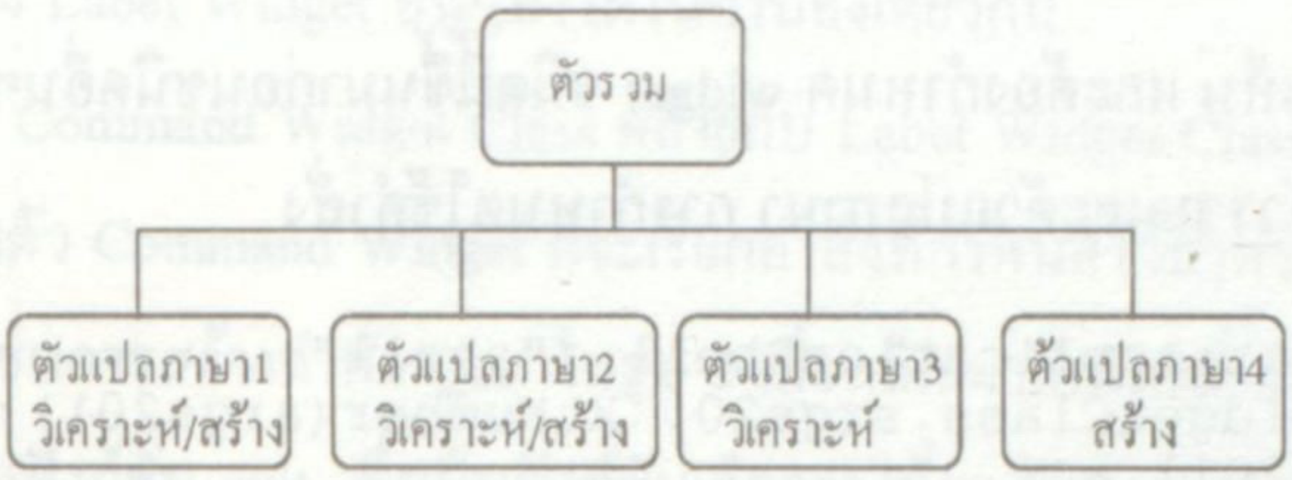
**2. ระบบรวม (Integrated System)** ระบบรวมเป็นโปรแกรมที่สร้างขึ้นเพื่อควบคุมกระบวนการประมวลผลของระบบการแปลภาษาโดยแต่ละเครื่องทำการประมวลผลแต่ละภาษาที่แตกต่างกัน โครงสร้างของระบบแบ่งเป็นสองส่วนคือ ตัวแปลภาษา (Translator) และตัวรวม (Integrator) ดังรูปที่ 1

ตัวแปลภาษาของแต่ละภาษาจะทำหน้าที่วิเคราะห์ภาษาและสร้างภาษา ตัวรวมทำหน้าที่แสดงผลหน้า ส่วนการรับ-ส่งข้อมูลเป็นการรับ-ส่งข้อมูลระหว่างตัวแปลภาษากับตัวรวม




รูปที่ 1 ระบบรวม

การเรียกใช้งานระบบรวม ระบบจะเริ่มงานที่ตัวรวมก่อนแล้วตัวรวมจะรอรับการติดต่อกับตัวแปลภาษา ตัวแปลภาษานี้จะรันกี่ตัวก็ได้แล้วแต่ต้องการ และในการเรียกใช้ตัวแปลภาษานั้นสามารถระบุได้ด้วยว่าต้องการให้ตัวแปลภาษาทำการวิเคราะห์ภาษา หรือสร้างภาษาอย่างใดอย่างหนึ่งหรือทั้งสองอย่าง เช่น เมื่อส่งอินพุทเป็นภาษา1 แล้วสั่งให้ระบบทำการแปล ระบบจะแปลประโยคอินพุทนั้นเป็นภาษา 2 และ 4 ซึ่งเราได้ตัวสร้างภาษาไว้ ส่วนตัวแปล 3 เรียกใช้งานเฉพาะตัววิเคราะห์ภาษาเท่านั้น ระบบจึงไม่ได้แปลเป็น 3




รูปที่ 2 ตัวอย่างการเรียกใช้งานระบบ

**2. การออกแบบระบบ** ในการสร้างระบบรวม แบ่งงานออกเป็น 2 ส่วนใหญ่ๆคือ การสร้างหน้าต่างแสดงผล ซึ่งในที่นี้ใช้ X Toolkit library ช่วยในการสร้างและการสร้างโปรแกรมควบคุมระบบตัวแปลภาษาซึ่งเขียนด้วยภาษาซีบนยูนิกซ์

**2.1 การสร้างหน้าต่างแสดงผล** ในที่นี้ใช้ X Toolkit ช่วยในการสร้างหน้าต่างแสดงผล X Toolkit Library เป็นไลบรารีสำเร็จรูปทีช่วย้การเขียนโปรแกรี่บน X Window System เนื่องจากฟังก์ชันใน X Toolkit Library มีลักษณะเป็น Object Oriented คือมองสิ่งต่างๆ เป็นก้อนวัตถุ โดยไม่จำเป็นต้องสนใจรายละเอียดภายใน

ลักษณะของ X Window System คือซอฟท์แวร์ที่ทำงานภายใต้ระบบปฏิบัติการ UNIX มีการติดต่อกับผู้ใช้ในลักษณะ Graphic User Interface (GUI) ปกติการพัฒนาโปรแกรมที่ทำงานภายใต้ X Window System จะใช้ภาษา C โดยใช้คำสั่งในไลบรารีมาตรฐาน การพัฒนาโปรแกรมชุดนี้ใช้ X Toolkit ซึ่งเป็น X Toolkit Library ในการพัฒนาโปรแกรม ทั้งการจัดหน้าต่างแสดงผล การจัดการสตริงข้อความ การรับ-ส่งอินพุต/เอาท์พุตจากผู้ใช้โดยใช้คำสั่งเพียงไม่กี่คำสั่งเท่านั้น

เนื่องจาก X Toolkit Library มีการใช้งานแบบ Object Oriented ซึ่งแบ่ง Object เป็น Class ต่างๆ ตามชนิดและการใช้งานของ Object นั้นๆ (ในที่นี้จะเรียกว่า Widget Class) ในการสร้างหน้าต่างแสดงผลของตัวรวมและตัวแปลภาษานี้จะ้สร้างบน widget class ต่างๆดังนี้

1. Shell Widget Class ทำหน้าที่เป็น parent widget ให้กับ Class อื่นๆและติดต่อกับ window manager เช่นถ้ามีการลดขนาด Shell Widget ผ่าน window manager ก็จะทำให้ child widget ที่อยู่ในตัวมันลดขนาดตามไปด้วย ปกติ widget ชนิดนี้เป็นพื้น

สี่เหลี่ยมผู้ใช้จะมองไม่เห็น และต้องกำหนด widget ชนิดนี้ขึ้นมาก่อนชนิดอื่นๆ ทั้งในหน้าต่างแสดงผลของตัวรวมและตัวแปลภาษา การกำหนดใช้คำสั่ง

```
shell  =  XtCreateApplicationShell ("shell",
transientShellWidgetClass,args20, XtNumber(args20));
static Arg Args20[] = {
        {XtNx, 7},
        {XtNy, 7},
        {XtNwidth, 1255},
        {XtNheight, 1010}};
```

2. Box Widget Class เป็นพื้นที่สี่เหลี่ยม มี resource ต่างๆให้เลือกใช้ เช่น การกำหนดสีพื้น ขนาดและสีของ border และอื่นๆ รวมทั้งคอยจัดการขนาดของ child widget ที่อยู่ในตัวมันด้วย จะต้องสร้าง Box Widgetบน Shell Widget ในตัวรวมสร้าง Box Widget 3 widget คือ boxshell เป็นพื้น boxshell1 เป็น box ที่อยู่ทางซ้าย และ boxshell2 อยู่ทางขวา โดยสร้าง boxshell บน shell ทั้ง boxshell1 และ boxshell2 บน boxshell ด้วยคำสั่ง

```
boxshell  =  XtCreateManagedWidget("boxshell",
 boxWidgetClass,  shell,NULL, 0);
boxshell1  =  XtCreateManagedWidget("boxshell1",
boxWidgetClass,boxshell,args98, XtNumber(args98));
static Arg args98[] = {
        {XtNwidth, 610},
        {XtNheight, 920}};
boxshell2  =  XtCreateManagedWidget("boxshell2",
boxWidgetClass,boxshell,args99, XtNumber(args99));
static Arg args99[] = {
        {XtNwidth, 610},
        {XtNheight, 920}};
```

ในทำนองเดียวกันก็ได้สร้าง Box Widget ในตัวแปลภาษา 3 widget เช่นกัน

3. Label Widget Class เป็นพื้นที่สี่เหลี่ยมมีสตริงข้อความไม่เกิน 1 บรรทัดอยู่ภายใน ซึ่ง winget นี้จะใช้เป็นชื่อเรื่อง การสร้าง Label Widget บน Box Widget เช่น IL-Display, T-Window, C-Window, E-Window และ J-Window ในรูป 3 และ IL-display กับ Output Display ในรูป 4 ทำโดยใช้คำสั่ง

```
Linterlingua = XtCreateManageWidget("Linterlingua",
 LabelWidgetClass, boxshell1, NULL, 0);
```

การสร้าง Label Widget อื่นๆทำได้ในทำนองเดียวกัน

4. Command Widget Class คล้ายกับ Label Widget Class ต่างกันที่เมื่อผู้ใช้คลิกเมาส์บนตัว Command Widget ก็จะเรียกคำสั่งที่กำหนดไว้สำหรับ Command Widget นั้นได้ เช่นการสร้างคำสั่ง Quit ในรูป 3 เมื่อคลิกเมาส์บนหน้าต่างของ Quit โปรแกรมก็จะเรียกฟังก์ชัน quit ซึ่งเป็นฟังก์ชันเลิกการใช้งานตัวรวม การสร้างใช้คำสั่ง

```
quit  = XtCreateManagedWidget("quit",
 commandWidgetClass,  boxshell1,NULL, 0);
XtAddcallback(quit, XtNcallback, Quit0, NULL);
```

ส่วนในตัวแปลภาษามี Command Widget ต่างๆดังต่อไปนี้ รูป 4

Load - เป็นคำสั่งที่ใช้เรียกไฟล์ข้อมูลภาษาต้นฉบับ แสดงผลในหน้าต่าง (1)

Save - เป็นคำสั่งที่ใช้บันทึกข้อมูลที่ปรากฏใน หน้าต่าง (1) ลงในไฟล์ข้อมูล

Clear - เป็นคำสั่งที่ใช้ลบข้อมูลในหน้าต่าง (1) และหน้าต่าง (2) ออกจากจอ

WCut - เป็นคำสั่งที่ใช้ตัดคำในประโยคภาษาต้นฉบับที่ต้องการแปล

Trans - เป็นคำสั่งที่ใช้ประมวลผลการแปลภาษา

Quit - เป็นคำสั่งให้ออกจากระบบ

ฟังก์ชันในส่วนของ Output Display

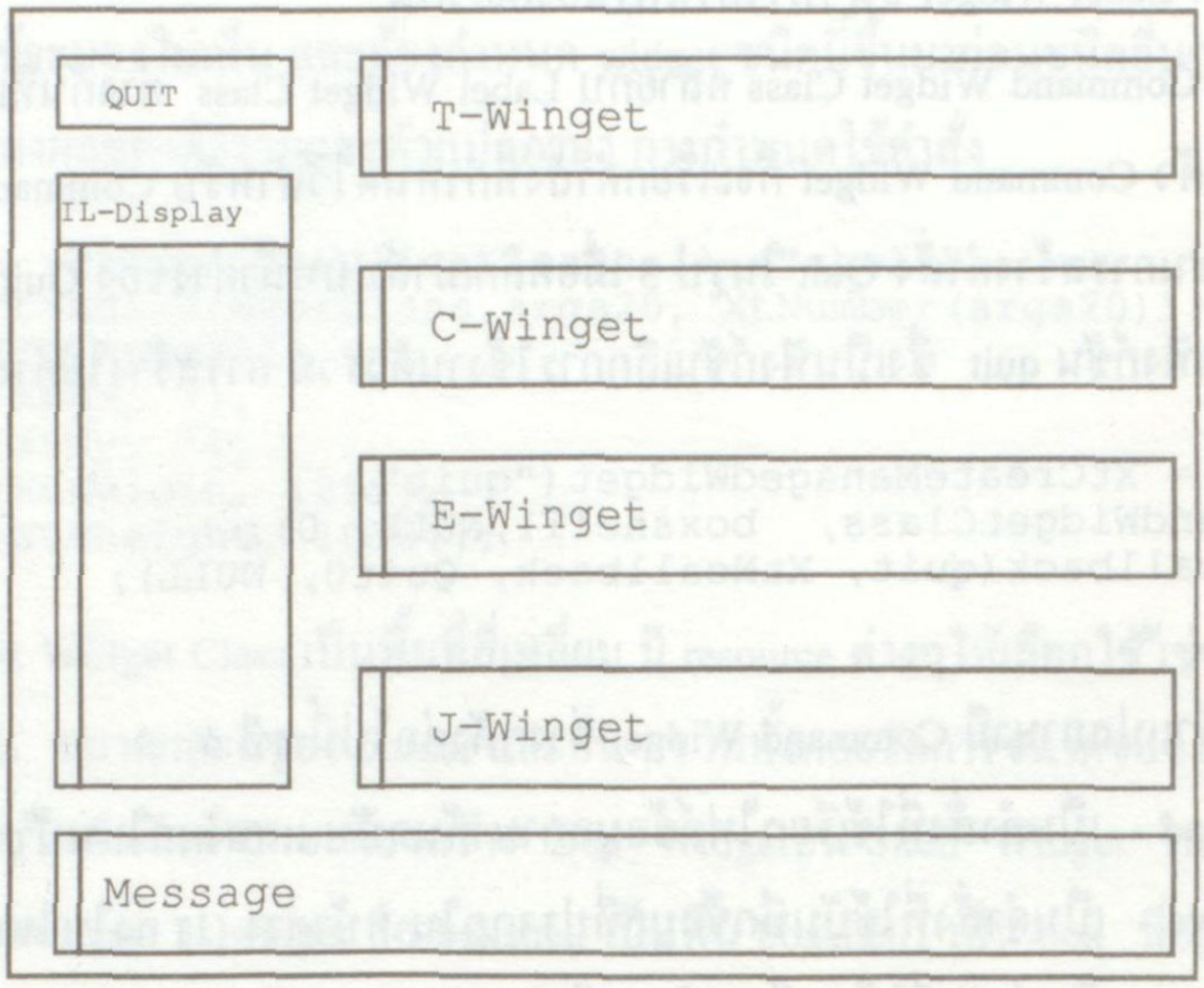
Save - เป็นคำสั่งที่ใช้บันทึกข้อมูลที่ปรากฏในหน้าต่าง (3) ลงในเอาต์พุตไฟล์

Clear - เป็นคำสั่งที่ใช้ลบข้อมูลที่ปรากฏในหน้าต่าง (3) ออกจากหน้าจอ

5. Dialog Widget Class ก็คล้ายกับ Label Widget แต่สามารถรับอินพุทจากผู้ใช้ได้ เช่น รับชื่อไฟล์ข้อมูล ในหน้าต่างตัวแปลภาษา คือ Filename ในรูป 4 ใน widget นี้แบ่งเป็น 2 ส่วนคือ ส่วนที่เป็นสตริง "Filename" จะทำหน้าที่รับข้อความอินพุทจากผู้ใช้ widget ชนิดนี้จะสร้างบน Box Widget ด้วยคำสั่ง

```
Filename    =    XtCreateManagedWidget("Filename",
   dialogWidgetClass,boxshell1, args99,
XtNumber(args99));
static Arg args99[] = {
        {XtNwidth, twidth},
        {XtNlabel, (XtArgVal)"Filename"};
        {XtNvalue, (XtArgVal)""},
```

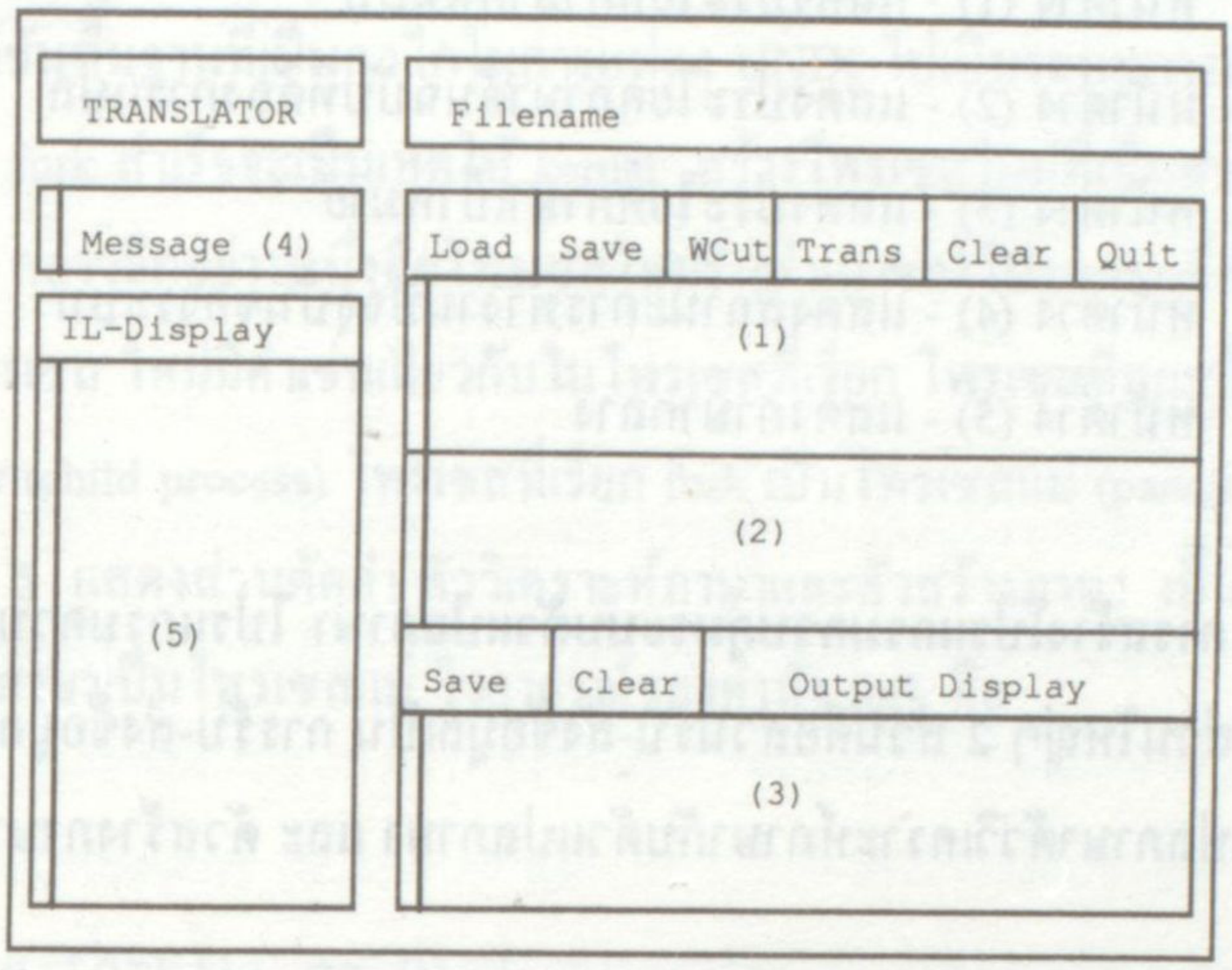
{NULL}};




รูปที่ 3 หน้าจอของตัวรวม

6. AsciiString Widget Class (Text Widget Class) เป็นพื้นที่ที่ใช้แสดงข้อความที่มากกว่า 1 บรรทัด เช่น ต้องการอ่านไฟล์ข้อมูลมาแสดงที่หน้าต่าง (1) รูป 4 AsciiString Widget มี resource ต่างๆมากมาย เช่นให้แก้ไขข้อมูลในหน้าต่างได้ซึ่งปกติจะให้อ่านได้อย่างเดียวเท่านั้น, กำหนดให้มี Scroll bar, กำหนดโหมดการเขียนว่าจะเป็น Append, Edit หรือ Read และรวมถึงการกำหนดขนาดของบัฟเฟอร์ด้วย AsciiString Widget จะต้องสร้างบน Box Widget ตัวอย่างการใช้คำสั่ง

```
interlingua = XtCreateManagedWidget("interlingua",AsciiSt
ringWidgetClass,
                boxshell1, args21, XtNumber(args21));
static Arg args21[] = {
        {XtNwidth, 600},
        {XtNheight, 818},
        {XtNscrollHorizontal, XawtextScrollWhenNeeded},
        {XtNscrollVertical, XawtextScrollAlways},
        {XtNeditType, XawtextEdit},
        {XtNstring, (XtArgVal)Interlingua},
        {XtNlength, BUFSIZE},
        {XtNtype, XawAsciiString}};
```




รูปที่ 4 หน้าจอของตัวแปลภาษา

หน้าต่างแสดงผลของตัวรวม

T-Window - แสดงประโยคภาษาไทย

C-Window - แสดงประโยคภาษาจีน

E-Window - แสดงประโยคภาษาอังกฤษ

J-Window - แสดงประโยคภาษาญี่ปุ่น

IL-Window - แสดงภาษากลาง

Message Window - แสดงสถานะการทำงานปัจจุบันของระบบ

หน้าต่างของตัวแปลภาษา

หน้าต่าง (1) - แสดงประโยคภาษาต้นฉบับ

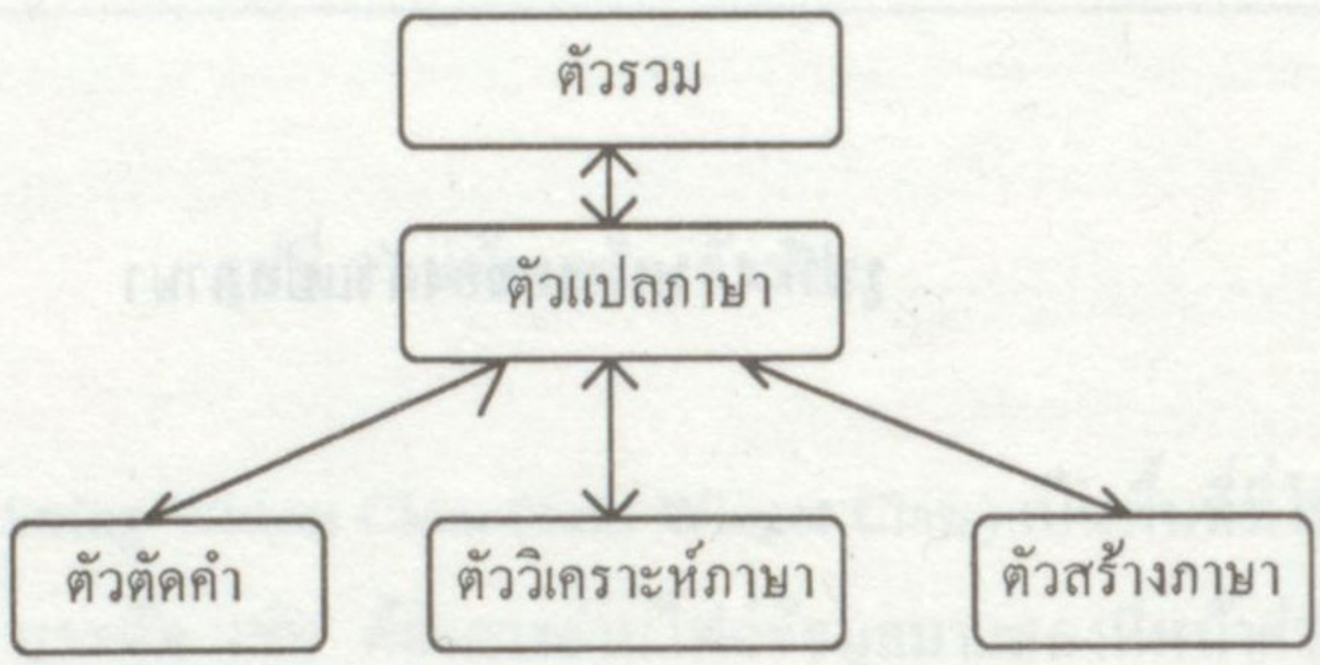
หน้าต่าง (2) - แสดงประโยคภาษาต้นฉบับที่ต้องการแปล

หน้าต่าง (3) - แสดงประโยคภาษาเป้าหมาย

หน้าต่าง (4) - แสดงสถานะการทำงานปัจจุบันของระบบ

หน้าต่าง (5) - แสดงภาษากลาง

**2.2 การสร้างโปรแกรมควบคุมระบบตัวแปลภาษา** โปรแกรมควบคุมระบบแบ่งออกเป็นส่วนใหญ่ๆ 2 ส่วนคือส่วนรับ-ส่งข้อมูลเป็น การรับ-ส่งข้อมูลระหว่างตัวรวมกับตัวแปลภาษาตัววิเคราะห์ภาษากับตัวแปลภาษา และ ตัวสร้างภาษากับตัว




รูป 5 การประมวลผลของระบบรวม

แปลภาษาโดยใช้ socket ดังแสดงในรูป 5 และส่วนประมวลผล ในส่วนนี้เป็นการสร้างหน้าต่างแสดงผล การตัดคำ การวิเคราะห์ภาษาและการสร้างภาษาการรันโพรเซสเหล่านี้ใช้ fork

**3. การประมวลผล** ดังที่กล่าวมาแล้วข้างต้นว่าในตัวแปลภาษาจะมีตัวประมวลผลหลัก คือ ตัวควบคุมหน้าต่างแสดงผล การตัดคำ ตัววิเคราะห์ภาษาและตัวสร้างภาษา การรันโพรเซสเหล่านี้เป็นการรันขนานกัน ซึ่ง UNIX เป็นระบบการทำงานแบบพหุคูณ (multi-tasking system) สามารถรันโพรเซสได้มากกว่าหนึ่งโพรเซสพร้อมกัน และ

UNIX ยังมี system call ที่มีประโยชน์ในการสร้างโพรเซสและการนำไปใช้อีกด้วย นั่นคือการใช้คำ สั่ง fork, exec, wait และ exit

fork เป็นฟังก์ชันพื้นฐานที่เป็นกลไกในการแปลง UNIX ไปเป็นระบบการทำงานแบบพหุคูณ ถ้าการ fork สำเร็จจะเป็นเหตุให้ kernal สร้างโพรเซสใหม่ที่เป็นสำเนาของโพรเซสที่เรียก กล่าวอีกอย่างหนึ่งคือโพรเซสใหม่รันสำเนาของโปรแกรมที่เป็นตัวสร้าง ตัวแปรในโพรเซส ใหม่มีค่าเช่นเดียวกับในโพรเซสที่เรียก โพรเซสที่ถูกสร้างขึ้นใหม่เรียกโพรเซสลูก(child process) โพรเซสที่เรียก fork เป็นโพรเซสแม่ (parent process) ในที่นี้ รูป 5 แสดงส่วนตัดคำ ตัววิเคราะห์ภาษาและตัวสร้างภาษา เป็นโพรเซสลูกส่วนตัวแปลภาษาเป็นโพรเซสแม่ ไวยากรณ์ของคำสั่ง fork คือ

```
int rtn, pid;
    ...
    if ((pid = fork()) == 0) {
        ....
    }
```

ฟังก์ชัน fork ไม่มีอาร์กิวเมนต์และคืนค่าเป็นจำนวนเต็ม pid ค่าของ pid จะเป็นการแยกโพรเซสแม่กับโพรเซสลูก ในโพรเซสแม่ ค่าของ pid จะเป็นจำนวนเต็มบวกที่ไม่ใช่ศูนย์ในขณะที่จะมีค่าเป็นศูนย์ในโพรเซสลูก เพราะว่าค่าที่คืนมาของโพรเซสแม่และลูกนั้นแตกต่างกัน เราจึงสามารถระบุการกระทำที่แตกต่างกันได้สำหรับโพรเซส ทั้งสองค่าที่คืนให้กับโพรเซสแม่ในตัวแปร pid เรียกว่า process-id ของโพรเซสลูกทุกโพรเซสที่เกิดจากการ fork จะมี process-id ของตัวเอง ซึ่งจะมีเพียงค่าเดียวเท่านั้นสำหรับแต่ละโพรเซสในช่วงเวลาหนึ่งๆ

ประโยชน์ของการใช้ fork คือสามารถใช้คู่กับ facilities อื่นๆของ UNIX ได้ ในที่นี้ใช้คู่กับคำสั่ง exec ซึ่งเป็นคำสั่งให้รันโปรแกรมใหม่ คำสั่ง execlp เป็นหนึ่งในชุดคำสั่ง exec จากตัวอย่างข้างล่างโปรแกรมใหม่คือโปรแกรมตัดคำในการสั่งรันตัววิเคราะห์ ภาษาและตัวสร้างภาษาก็ทำได้ในทำนองเดียวกัน

จากตัวอย่างเป็นการใช้ fork คู่กับ exec การ fork เป็นการรันโปรแกรมตัดคำในโพรเซสลูก ขณะนี้โพรเซสแม่จะหยุดรอให้โพรเซสลูกทำงานเสร็จสิ้น ก่อนด้วยคำสั่ง wait การทำงานแบ่งออกเป็น 3 ส่วน ก่อน fork หลัง fork และหลัง execlp

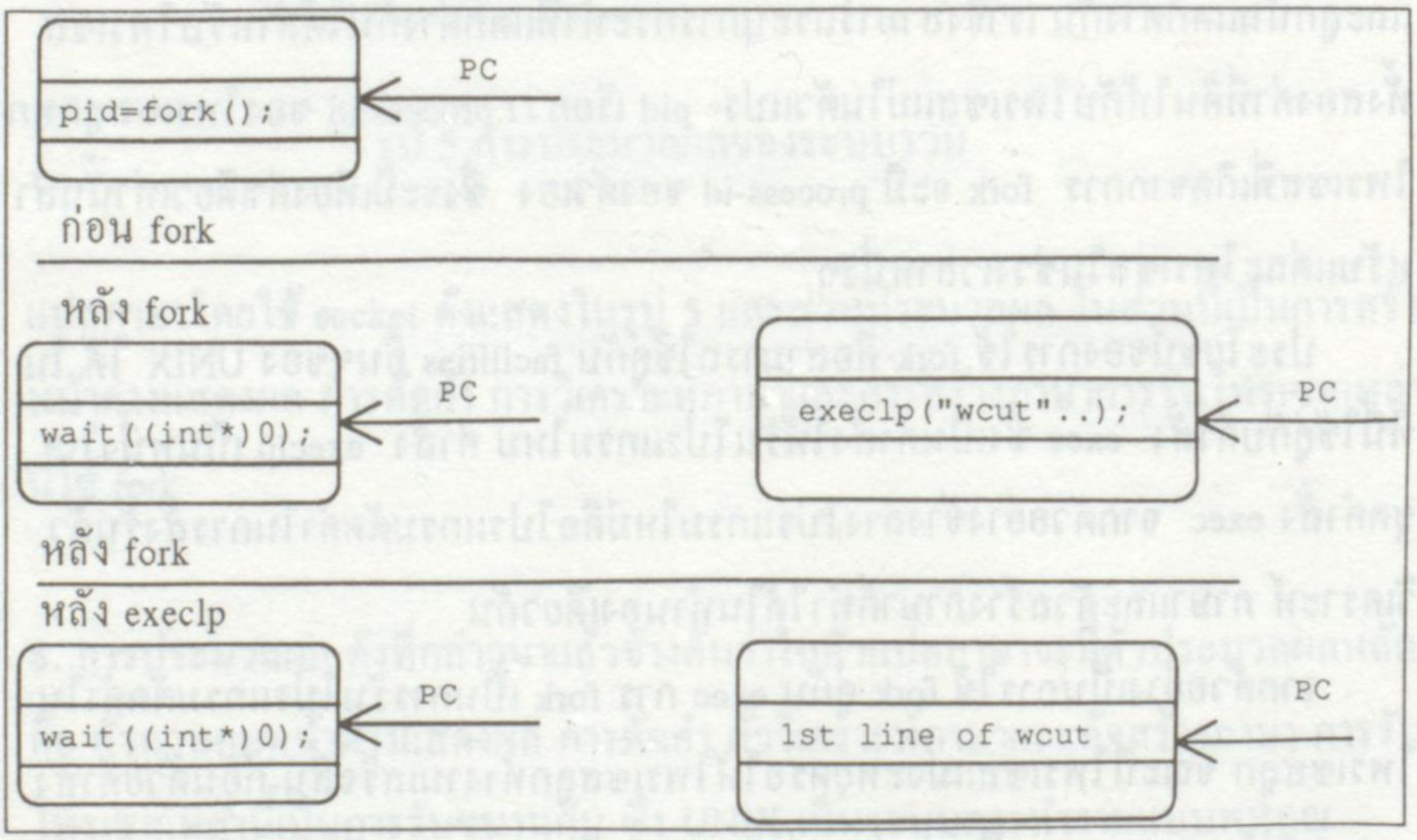
```
int rtn, pid;
    ...
```

```
    if ((pid = fork()) == 0) {
       chmod(Filename1, 00600);
       rtn = execlp("wcut", client_data, BFile,Filename1,
ADIC,0);
       if (rtn == -1) {
          sprintf(Message, "Execution : Error at '%s'\n",
client_data);
          Outmessage(Message);
          exit(0);
       } }
    if (pid > 0) {
       wait((int *)0);
       Outmessage("Word cut : success\n");
       ... }
```

เริ่มพิจารณาจาก ก่อน fork (รูป 6) ขณะนี้เป็นการทำงานของโพรเซส A และ program counter ก็ชี้ไปที่คำสั่ง fork เป็นการบอกว่าเป็นคำสั่งถัดไปที่จะทำการ

หลังจากเรียก fork จะมี 2 โพรเซส (A และ B) โพรเซสแม่ A กำลัง execute wait ทำให้โพรเซสแม่พักไว้ก่อนจนกว่าโพรเซส B จะสิ้นสุด ในระหว่างนั้น B ใช้ execlp โหลดโปรแกรมตัดคำ

ในส่วนหลัง execlp ขณะนี้เป็นการรันโปรแกรม wcut program counter สำหรับ B จะชี้ไปที่คำสั่งแรกของ wcut เพราะว่า A คอยให้ B เสร็จสิ้น program counter ของ A จึงไม่เปลี่ยนตำแหน่ง




รูป 6 การใช้ fork ร่วมกับ execlp

**4. การรับ-ส่งข้อมูลของระบบ** จากรูป 2 เริ่มจากตัวแปลภาษา(1) รับอินพุตซึ่งเป็นประโยคภาษาที่ต้องการ แปลจากหน้าจอ (รูป 3) แล้วส่งประโยคนี้ไปแสดงผลบนหน้าจอ (รูป 4) ของตัวรวม จากนั้นตัวแปลภาษา (1) จะทำการวิเคราะห์ประโยคภาษาต้น ฉบับให้เป็น ภาษากลาง(Interlingua) ด้วยตัววิเคราะห์ภาษา แล้วส่งภาษากลางนี้ไปที่ตัวรวม ตัว รวมจะแสดงภาษากลางออกหน้าจอ และส่งภาษากลางนี้ไปยังเครื่องที่เป็นตัว แปลภาษา(1), (2)และ(4) ตัวแปลภาษา (1), (2)และ(4)รับภาษากลางนั้นแล้วสร้างภาษา โดยใช้ตัวสร้างภาษา จากนั้นแต่ละเครื่องจะแสดงผลในหน้าจอของตัวเอง พร้อมทั้งส่งกลับไปยังตัวรวม ตัวรวมแสดงผลหน้าจอ การรับ-ส่งข้อมูลในที่นี้เป็นการส่งไฟล์ข้อความ

การติดต่อระหว่างโพรเซสและการติดต่อระหว่างระบบในที่นี้ใช้ sockets ซึ่งเป็นแบบหนึ่งของการติดต่อระหว่างโพรเซส(interprocess communication(IPC))ของระบบ UNIX การติดต่อเป็นแบบ client/server ในที่นี้ ตัวรวมเป็น server และตัวแปลภาษาเป็น client ในที่นี้เปิด socket ระหว่างตัวรวมกับตัวแปลภาษา ตัววิเคราะห์ภาษา กับตัวแปลภาษาและตัวสร้างภาษากับตัวแปลภาษา การเปิด socket มีไวยากรณ์ดังนี้

```
server_socket-open()
{
   int is_id;
   if ((so_is = socket(AF_INET, SOCK_STREAM, ))) < 0)
      Error("socket");
   if (gethostname(hostname, sizeof(hostname)) < 0)
      Error("gethostname");
   if ((hostentp = gethostbyname(hostname)) == 0)
      Error("gethostbyname");
   name_is.sin_family = AF_INET;
   name_is.sin_port = so_portr_is;
   name_is.sin_addr.s_addr = *((int *) (hostentp-
>h_addr));
   namelen = sizeof(name_is);
   if (bind(so_is, &name_is, namelen) < 0) Error("bind");
   fprintf(stderr, "Socket : Waiting for connection at
%d\n", so_is);
   if (listen(so_is,5)) Error("listen");
}
```

## 5. ขั้นตอนการแปลภาษา

1. การเรียกตัวรวม ใช้คำสั่ง

KMITT > INT

จะแสดงผลเป็นรูป 3

2. การเรียกตัวแปลภาษา ใช้คำสั่ง

KMITT > tran_ int_host option

tran_ - ชื่อโปรแกรมการแปลภาษาที่ต้องการ

tranT โปรแกรมแปลภาษาไทย

tranC โปรแกรมแปลภาษาจีน

tranJ โปรแกรมแปลภาษาญี่ปุ่น

tranE โปรแกรมแปลภาษาอังกฤษ

int_host - hostname ของเครื่องที่ใช้เป็นตัวรวม

option A เป็นตัววิเคราะห์ภาษา

G เป็นตัวสร้างภาษา

- เป็นทั้งตัววิเคราะห์และสร้างภาษา

3. การแปลข้อความที่ต้องการ

ตัวแปลภาษารับอินพุตซึ่งเป็นข้อความที่ต้องการแปล (หน้าต่างรูป 4) โดย

3.1 ถ้าต้องการเลือกข้อความจากไฟล์ข้อมูลที่มีอยู่แล้ว

- ใส่ชื่อไฟล์ข้อมูลลงในหน้าต่าง " Filename"
- ใช้คำสั่ง Load ไฟล์ข้อมูลจะปรากฏในหน้าต่าง (1)
- เลือกข้อความที่ต้องการจากหน้าต่าง (1) โดยใช้เมาส์ copy มาที่หน้าต่าง (2) หรือจะป้อนข้อมูลนั้นจากแป้นพิมพ์ก็ได้
- ใช้คำสั่ง Trans เพื่อแปลข้อความ
- ผลจากการแปลจะปรากฏในหน้าจอของตัวรวม และในหน้าต่าง (3) ของตัวแปลภาษาอื่นๆที่ทำหน้าที่สร้างภาษา

3.2 ถ้าต้องการป้อนข้อความที่ต้องการแปลจากหน้าแป้นพิมพ์

- ให้ป้อนข้อความลงในหน้าต่าง (2)
- ใช้คำสั่ง Trans เพื่อแปลข้อความ
- ผลจากการแปลจะปรากฏในหน้าจอของตัวรวม และในหน้าต่าง(3) ของ

**6. ข้อสรุป** การสร้างระบบรวมในที่นี้ได้ทดลองใช้กับภาษาไทย อังกฤษ ญี่ปุ่น และจีน โดยทำการ แปลประโยคตัวอย่างจากภาษาหนึ่งไปเป็นภาษาอื่นๆ ผลที่ได้เป็นที่น่าพอใจ

**เอกสารอ้างอิง**

1. รายงานการวิจัยและพัฒนา
โครงการพัฒนาระบบรวมเพื่อการแปลภาษาด้วยคอมพิวเตอร์
(ระยะที่ 2:2532)

2. Haviland Keith and Salama Ben, UNIX system programming. Addison-Wesley Publishing Company. 1987.

3. O'Reilly Tim and Langley Mark, X Toolkit Intrinsics Reference Manual. O'Reilly&Associates,Inc. 1990.

4. Nye Adrian and O'Reilly Tim, X Toolkit Intrinsics Programming Manual. O'Reilly&Associates,Inc. 1990.

# อักขรวิธีภาษาไทยสำหรับระบบX Window

# Thai Input/Output Method for X Window System

วิรัช ศรเลิศล้ำวาณิช

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้

ศูนย์เทคโนโลยีอิเลคทรอนิกส์และคอมพิวเตอร์แห่งชาติ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

**1. บทนำ** การวิจัยและพัฒนาอักขรวิธีภาษาไทยสำหรับใช้ในระบบ X Window เกิดขึ้นมาเนื่องจากความจำเป็นในการวิจัยระบบเครื่องแปลภาษา ซึ่งจำเป็นที่จะต้องคำนึงถึงการจัดการแบบหลากภาษาเป็นพื้นฐาน นอกจากจะต้องมีมาตรการที่ดีที่จะให้ระบบเป็นสากลได้แล้ว ยังจะต้องมีวิธีการที่ดีในการจัดเก็บข้อมูลเพื่อให้เอื้อต่อขั้นตอนต่างๆในการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ด้วย จะเห็นได้ว่าความจำเป็นหรือข้อจำกัดในการจัดระบบนั้นละเอียดอ่อนกว่าการวางระบบเพียงเพื่อการประมวลคำในระบบของ Word processor เป็นอย่างมาก และในที่สุดอักขรวิธีอันนี้ก็จะช่วยยกระดับการประมวลทางภาษาให้กับระบบของ Word processor ทั่วไปได้

**2. ความเป็นมาของการจัดระบบอักขระสำหรับภาษาไทย** การกำหนดรหัสไทยเพื่อใช้ในเครื่องคอมพิวเตอร์นั้นได้มีการพยายามทำกันมานานแล้ว เริ่มตั้งแต่ปีพ.ศ. 2510 และในปัจจุบันนี้หลังจากที่สำนักงานมาตรฐานอุตสาหกรรม (สมอ.) ได้ออกประกาศใช้ มอก.620-2529 (ดังปรากฏในรูปที่ 1) เป็นรหัสมาตรฐานขึ้น การจัดการกับภาษาไทยในระบบคอมพิวเตอร์ก็เริ่มมีการยึดถือเป็นหลักเดียวกัน ถึงแม้ว่าจะยังคงมีการใช้กันมากกว่า 1 ระบบก็ตาม แต่ทุกระบบก็มีข้อกำหนดที่ยึดถือกันอยู่ จึงทำให้สามารถมีการแลกเปลี่ยนข้อมูลหรือเปลี่ยนแปลงระบบของรหัสที่ใช้ไปมาระหว่างกันและกันได้

| 4 บิทต้น / 4 บิทท้าย | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | SP | 0 | @ | P | | p | | | | ฐ | ก | ะ | เ | ๐ |
| 1 | | | ! | 1 | A | Q | a | q | | | ก | ฑ | ม | ั | แ | ๑ |
| 2 | | | " | 2 | B | R | b | r | | | ข | ฒ | ย | า | โ | ๒ |
| 3 | | | # | 3 | C | S | c | s | | | ฃ | ณ | ร | ำ | ใ | ๓ |
| 4 | | | $ | 4 | D | T | d | t | | | ค | ด | ฤ | ิ | ไ | ๔ |
| 5 | | | % | 5 | E | U | e | u | | | ฅ | ต | ล | ี | ๅ | ๕ |
| 6 | | | & | 6 | F | V | f | v | | | ฆ | ถ | ฦ | ึ | ๆ | ๖ |
| 7 | | | ' | 7 | G | W | g | w | | | ง | ท | ว | ื | ็ | ๗ |
| 8 | | | ( | 8 | H | X | h | x | | | จ | ธ | ศ | ุ | ่ | ๘ |
| 9 | | | ) | 9 | I | Y | i | y | | | ฉ | น | ษ | ู | ้ | ๙ |
| A | | | * | : | J | Z | j | z | | | ช | บ | ส | ฺ | ๊ | ฯ |
| B | | | + | ; | K | [ | k | { | | | ซ | ป | ห | | ๋ | ฿ |
| C | | | , | < | L | \ | l | \| | | | ฌ | ผ | ฬ | | ์ | |
| D | | | - | = | M | ] | m | } | | | ญ | ฝ | อ | | ๎ | |
| E | | | . | > | N | ^ | n | ~ | | | ฎ | พ | ฮ | | ็ | |
| F | | | / | ? | O | _ | o | | | | ฏ | ฟ | ฯ | ฿ | ๏ | |

ISO 646-1983 มอก.620-2529

รูปที่ 1. ตารางรหัสชุดอักขระพื้นฐาน

การพัฒนาโปรแกรมสำหรับการจัดการข้อมูลที่ผ่านมา เช่น การพัฒนาโปรแกรมบรรณาธิกรก็ดี มักจะถูกออกแบบให้สามารถจัดการได้แบบ 2 ภาษา (Bilingual) คือ ภาษาอังกฤษ กับภาษาของตน ดังนั้นการกำหนดรหัสของภาษานั้นๆก็จะคำนึงถึงเฉพาะ 2 ภาษาดังกล่าวเท่านั้น ปัจจุบันการแลกเปลี่ยนข่าวสารหรือการพัฒนาระบบเริ่มเป็นสากลมากขึ้น โดยเฉพาะระบบ UNIX ซึ่งออกแบบไว้ให้เป็นระบบ Multi user ที่สามารถทำงานได้ในระบบเครือข่าย เริ่มมีการใช้กันอย่างแพร่หลาย และเมื่อคนต่างชาติต่างภาษาต้องการที่จะติดต่อหรือแลกเปลี่ยนข้อมูลที่เป็นบทความกัน หรือใช้คอมพิวเตอร์ในสถานภาพที่เป็นของภาษาตนเองแล้ว ระบบคอมพิวเตอร์ที่ใช้อยู่นั้นก็ควรจะสามารถจัดการกับข้อมูลในภาษานั้นๆได้

เมื่อการจัดการกับข้อมูลที่มีหลากภาษา (Multi-lingual text) เริ่มมีความจำเป็นมากขึ้น ดังนั้นกลุ่มนักวิจัยและพัฒนาจากหลายฝ่ายจึงมีความตื่นตัวในการที่จะพัฒนาให้

ระบบพื้นฐานสามารถที่จะรับและจัดการกับรหัสของหลายๆภาษาได้ในเวลาเดียวกัน ปัจจุบันนี้ได้มีอยู่ 3 องค์กรใหญ่ๆที่พยายามจะจัดระบบอักขระให้กับทุกๆภาษาที่มีใช้กันอยู่ องค์กรทั้ง 3 คือ Unicode Consortium, MIT X Consortium และ ISO (the International Organization for Standardization)

**3. รหัส EUC (Extended UNIX Code)** บริษัท AT&T แห่งประเทศสหรัฐอเมริกาได้กำหนดรหัสเสริมนี้ขึ้นสำหรับใช้ในระบบ UNIX เพื่อการแสดงผลของภาษาต่างๆ รหัสภายใน(Internal Code) ได้กำหนดไว้ดังในรูปที่ 2

| Code Set | Internal Code |
|---|---|
| Set 0 (ASCII) | 0XXXXXXX |
| Set 1 | 1XXXXXXX [1XXXXXXX [...]] |
| Set 2 | SS2 1XXXXXXX [1XXXXXXX [...]] |
| Set 3 | SS3 1XXXXXXX [1XXXXXXX [...]] |

SS2 : 0x8e

SS3 : 0x8f

รูปที่ 2. ตารางแสดงรหัสเสริมของ UNIX (EUC)

รหัสนี้กำหนดขึ้นเพื่อเป็นมาตรฐานสำหรับใช้ในการติดต่อกันทั้งในลักษณะของไฟล์ในจานแม่เหล็ก เทปแม่เหล็ก เป็นต้น หรือการส่งข้อมูลกันระหว่างกระบวนการ (process) System call interface เป็นต้น

รหัสEUC นี้ได้นำมาใช้เป็นมาตรฐานในระบบUNIX System V Release 4 (SVR4) MNLS (Multi National Language Supplement) ด้วย ซึ่งใน SVR4 ก็จะมี Library ต่างๆเพิ่มขึ้นมาเพื่อจัดการกับตัวอักษรประเภท Multi byte

**4. รหัส EUC สำหรับอักขระไทย** อักขระไทยที่มีใช้อยู่ในปัจจุบันตามมาตรฐานของ มอก.620-2529 สามารถแบ่งออกได้เป็น 5 ประเภทตามลักษณะของการใช้ กล่าวคือ **พยัญชนะ สระ วรรณยุกต์ ตัวเลข** และเครื่องหมายพิเศษประกอบการเขียน

การประสมอักษรก็ยังมีลักษณะพิเศษโดยที่ว่าไม่ได้เรียงตามลำดับของการผสมของหน่วยเสียงเสมอไป แต่ตำแหน่งของการปรากฏของอักขระแต่ละตัวจะมีตำแหน่งที่ตายตัวเมื่อเทียบกับตัวพยัญชนะที่ถูกประสม อาทิเช่น สระ"เ" สระ"แ" หรือ สระ"โ" จะปรากฏอยู่ที่ตำแหน่งหน้า สระ"ะ" สระ"า" จะปรากฏอยู่ที่ตำแหน่งหลัง สระ"ุ" สระ"ู" จะปรากฏอยู่ที่ตำแหน่งล่าง สระ"ิ" สระ "ี" หรือ วรรณยุกต์ทุกรูปจะปรากฏอยู่ที่ตำแหน่งบน เป็นต้น ลักษณะเช่นนี้ทำให้คอมพิวเตอร์ต้องมีฟังก์ชั่นพิเศษมาช่วยในการจัดการทั้งในด้านการป้อนเข้า(input) การแสดงผลบนจอ(output) และการประมวลผลทางภาษาเช่น การจัดเรียงคำ การแบ่งคำ การสืบค้น เป็นต้น

จากลักษณะดังกล่าวข้างบนทำให้ภาษาไทยมีบางกลุ่มของอักขระที่เป็น Non-spacing Character จึงทำให้จำนวนตัวอักขระต่อ 1 เซลไม่เท่ากันในแต่ละเซล และทำให้เซลที่มีการประสมของอักขระที่เหมือนกันอาจมีการเก็บลำดับของรหัสอักขระที่แตกต่างกันออกไปได้ซึ่งจะมีผลทำให้เกิดข้อแตกต่างระหว่างข้อมูลกับผลที่ปรากฏให้เห็นบนจอ

เพื่อที่จะแก้ปัญหาดังกล่าว และลดภาระของซอฟต์แวร์ที่จะต้องมาจัดการกับผลที่มีข้อขัดแย้งที่อาจเกิดขึ้นได้ ดังนั้นจึงได้พิจารณาการเก็บอักขระแบบ 2 ไบท์ต่อ 1 เซล เพื่อที่จะให้แต่ละหน่วยของรหัสสามารถแทนอักขระที่ประสมแล้วของแต่ละเซลได้ ชุดรหัส (Code set) ของEUC ที่เลือกใช้จึงมี Code set 0 ซึ่งเป็นรหัสของ ASCII พื้นฐาน แล้วกำหนด Code set 1 ที่มีความยาว 2 ไบท์ให้เป็นรหัสสำหรับอักขระไทย

**5. วิธีการเก็บอักขระไทยแบบ 2 ไบท์** การเก็บจะต้องคำนึงถึงความสะดวกในการจัดเก็บ การแสดงผล และที่สำคัญคือจะต้องสามารถแปลงไปมากันอย่างสะดวกกับรหัสมาตรฐาน (มอก.620-2529) ได้เป็นอย่างน้อย

เนื่องจากการแสดงผลบนจอจะเป็นดังในรูปที่ 3. ซึ่งในแต่ละเซลแม้ว่ามีอักขระปรากฏอยู่ได้ถึง 4 ตำแหน่งก็ตามแต่ในแต่ละเซลจะมีอักขระประสมกันอยู่ได้สูงสุดไม่เกิน 3 ตัวเท่านั้นเอง

| | | | | |
|---|---|---|---|---|
| | ◌ | | 1 | TOP |
| | ◌ | ◌ | 2 | ABOVE |
| Base line | ก | ค | 3 | BASE |
| | | ◌ | 4 | BELOW |

รูปที่ 3. แสดงตำแหน่งของอักขระในแต่ละเซล

เมื่อกำหนดให้ไบท์แรกเป็นส่วนที่เก็บรหัสของอักขระที่ปรากฏอยู่บน Base line โดยให้มีรหัสตามมาตรฐานมอก.620-2529 แล้วให้ไบท์ที่ 2 เป็นส่วนที่เก็บรหัสของอักขระที่ประสมอยู่หรือกลุ่มของอักขระที่เป็น Non-spacing Character ฉะนั้นรหัส 2 ไบท์แต่ละชุดที่นอกเหนือจาก Code set 0 จะสามารถแทนที่ตัวอักขระประสมภาษาไทยในแต่ละเซลได้

| 1 X X X X X X X | 1 X X X X X X X |
|---|---|
| ไบท์ที่ 1<br>(รหัสมอก.620-2529) | ไบท์ที่ 2<br>(รหัสอักขระประสม)<br>(Verbal-Tone Code) |

รูปที่ 4. แสดงชุดรหัส 2 ไบท์สำหรับอักขระไทยใน 1 เซล

**5.1 การกำหนดรหัสในไบท์ที่ 2** รหัส 2 ไบท์สำหรับอักขระไทยนี้เรียกว่า X-TIS (eXtended Thai Industrial Standard code set) ในไบท์ที่ 2 ยกเว้นบิทที่ MSB (Most Significant Bit) ซึ่งจำเป็นที่จะต้องเป็น 1 เสมอแล้วจะถูกแบ่งออกเป็น 2 กลุ่มโดยอาศัยกฎที่ว่าจะไม่มีการประสมสระ(หรือวรรณยุกต์)มากกว่า 1 ตัวกับพยัญชนะตัวหนึ่งๆ

ดังนั้นอักขระในไบท์ที่ 2 มีการจัดกลุ่มและกำหนดรหัสตามตารางต่อไปนี้

| | รูป | มอก. 620-2529 | รหัส V-T* | Bit array | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| กลุ่มสระ | - | - | 0x00 | 1 | 0 | 0 | 0 | 0 | X | X | X |
| | ◌ั | 0xd1 | 0x08 | 1 | 0 | 0 | 0 | 1 | X | X | X |
| | ◌ิ | 0xd4 | 0x10 | 1 | 0 | 0 | 1 | 0 | X | X | X |
| | ◌ี | 0xd5 | 0x18 | 1 | 0 | 0 | 1 | 1 | X | X | X |
| | ◌ึ | 0xd6 | 0x20 | 1 | 0 | 1 | 0 | 0 | X | X | X |
| | ◌ื | 0xd7 | 0x28 | 1 | 0 | 1 | 0 | 1 | X | X | X |
| | ◌ุ | 0xd8 | 0x30 | 1 | 0 | 1 | 1 | 0 | X | X | X |
| | ◌ู | 0xd9 | 0x38 | 1 | 0 | 1 | 1 | 1 | X | X | X |
| | ◌ฺ | 0xda | 0x40 | 1 | 1 | 0 | 0 | 0 | X | X | X |
| | ◌เ | 0xee | 0x48 | 1 | 1 | 0 | 0 | 1 | X | X | X |
| กลุ่มวรรณยุกต์ | - | - | 0x00 | 1 | X | X | X | X | 0 | 0 | 0 |
| | ◌็ | 0xe7 | 0x01 | 1 | X | X | X | X | 0 | 0 | 1 |
| | ◌่ | 0xe8 | 0x02 | 1 | X | X | X | X | 0 | 1 | 0 |
| | ◌้ | 0xe9 | 0x03 | 1 | X | X | X | X | 0 | 1 | 1 |
| | ◌๊ | 0xea | 0x04 | 1 | X | X | X | X | 1 | 0 | 0 |
| | ◌๋ | 0xeb | 0x05 | 1 | X | X | X | X | 1 | 0 | 1 |
| | ◌์ | 0xec | 0x06 | 1 | X | X | X | X | 1 | 1 | 0 |
| | ◌ | 0xed | 0x07 | 1 | X | X | X | X | 1 | 1 | 1 |

*รหัสV-T : Verbal-Tone Code คือรหัสภายในที่ใช้แทน Non-spacing Character ในไบท์ที่ 2

รหัสในไบท์ที่ 2 เริ่มตั้งแต่ 0xb0 เป็นต้นไปทั้งนี้เพื่อสงวนไว้สำหรับการกำหนด Control Code ต่างๆในภายหลัง

ตัวอย่างการประสมอักขระ

ก => 0xa1b0 + 0x00 + 0x00 => 0xa1b0

ก็ => 0xa1b0 + 0x00 + 0x01 => 0xa1b1

กี่ => 0xa1b0 + 0x18 + 0x02 => 0xa1ca

ชู้ => 0xaab0 + 0x38 + 0x03 => 0xaaeb

**5.2 ข้อดีของการใช้ชุดรหัส X-TIS** การเก็บอักขระในชุดรหัส X-TIS นี้นอกจากจะอยู่ในมาตรฐานของ EUC แล้วยังช่วยในการแก้ปัญหาการประมวลผลทางภาษาได้เป็นอย่างมาก ดังต่อไปนี้

เนื่องจากว่ารหัสของ Non-spacing Character ถูกเก็บรวมเข้ากับตัวพยัญชนะที่ถูกประสมและมีตำแหน่งในระบบรหัสที่สัมพันธ์กับการแสดงผลบนจอ ดังนั้น

ก. จะไม่มีรหัสของ Non-spacing Character ซ้ำกันเกิดขึ้นถึงแม้ว่าจะมีการป้อนซ้ำในที่เดิมก็ตาม ระบบรหัส X-TIS จะไม่ยอมให้เกิดกรณีของการป้อนข้อมูลของสายอักขระที่ผิด

ข. รหัสของ Non-spacing Character จะถูกจัดเข้าตามตำแหน่งที่ถูกต้องเสมอไม่ซึ่งทำให้ลำดับของการป้อนไม่มีผลแต่อย่างไรต่อคำที่มีตัวสะกดเดียวกัน

ค. โดยพื้นฐานแล้วการประสมของอักขระที่ถูกต้องตามอักขรวิธีเท่านั้นที่จะทำการป้อนเข้าไปได้

ง. ความยาวรหัสของอักขระไทยจะคงที่เสมอคือ 2 ไบท์ ถึงแม้ว่าจะทำให้ผลรวมของความยาวรหัสโดยเฉลี่ยของไฟล์บทความหนึ่งจะเพิ่มขึ้นก็ตาม แต่ความยาวรหัสของอักขระที่คงที่จะทำให้การประมวลผลต่างๆสามารถทำได้อย่างรวดเร็วและถูกต้องได้ โดยเฉพาะในการหาขอบเขตของของคำ

จ. ขจัดปัญหาที่เกิดจากการ Cut&Paste บนหน้าจอซึ่งผู้ใช้ไม่จำเป็นที่จะต้องระวังตำแหน่ง Non-spacing Character อีกต่อไป

ฉ. ขจัดปัญหาในการหา Insert point ใน text

**6. อักขรวิธีในระบบ X Window** งานวิจัยนี้ได้ทดลองทำบน X11R4 โดยทำการปรับปรุง Input/Output Method ของ Widget ที่ Project Athena ได้เผยแพร่ไว้ใน Library ของ Athena Widgets (libXaw.a) การปรับปรุงในครั้งนี้จึงทำให้ผู้พัฒนาโปรแกรมโดยใช้ Library ของ X Toolkit สามารถที่จะสร้าง Window ที่สามารถรับและแสดงผลภาษาไทยได้ทุกรูปแบบตามที่ Athena Widgets มีมาให้

ใน Mode ของการป้อนข้อมูลที่เป็นภาษาไทย Input/Output Method ใน Text Widget นอกจากจะ Map คีย์ให้เป็นไปตามมาตรฐานของ มอก.820-2531 แล้วยังทำการตรวจเช็คและแปลงโค้ดให้เป็นไปตามหลักการของ X-TIS ที่กล่าวไว้ในหัวข้อที่ 5 ผู้

พัฒนาโปรแกรมเพียงแต่เขียนโปรแกรมตามข้อกำหนดของ X Toolkit เท่านั้นโดยไม่ต้องคำนึงถึงการจัดการทางภาษาไทยแต่อย่างไร ก็จะได้ Window ที่สามารถรับและแสดงผลที่เป็นภาษาไทย

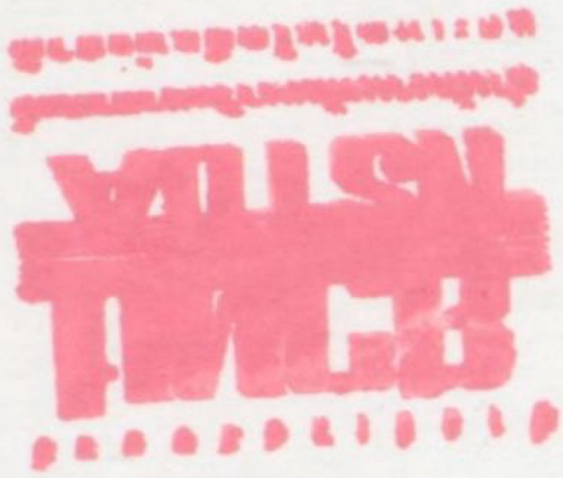
Font ที่เตรียมไว้เพื่อการแสดงผลบนจอชนิด Bitmap เป็นแบบ Fix-width ขนาด 12x24 point Font ชุดนี้สร้างเป็นแบบ Bitmap แล้วแปลงเป็น SNF Font ซึ่งเป็น format ที่เรียกใช้ได้ในระบบ X Window

**7. สรุป** ผลงานนี้ถึงแม้ว่าจะอยู่ในขั้นทดลองก็ตามแต่ผลของการใช้งานเป็นที่น่าพอใจอย่างยิ่ง ซึ่งสามารถเพิ่มประสิทธิภาพของผู้พัฒนาโปรแกรมบนระบบ X Window และที่สำคัญคือชุดรหัส X-TIS เป็นพื้นฐานที่ดีสำหรับการประมวลผลทางภาษา ไม่ว่าจะเป็นการหาขอบเขตของคำในกระบวนการตัดคำ (Word Segmentation) การจัดเก็บและการสืบหาคำในพจณานุกรมที่จำเป็นต้องอาศัยความถูกต้องทางตัวสะกดเป็นอย่างสูง หรือการค้นหาข้อความในโปรแกรม KWIC (Key-Word-In-Context) เป็นต้น

ดังได้กล่าวไว้ข้างต้น งานวิจัยนี้ได้ทดลองจริงบน X11R4 ซึ่งปัจจุบันนี้ release ใหม่ของ X Window เป็น X11R5 แล้ว และที่สำคัญที่สุดคือภาษาไทยยังไม่มีการกำหนดมาตรฐานของ Input/Output Method ในระบบของ X Window หวังอย่างยิ่งว่าในอนาคตอันใกล้นี้จะมีมาตรฐานที่เหมาะสมกับภาษาไทยจริงๆในมาตรฐานสากล

## เอกสารอ้างอิง


1. Adrian Nye and Tim O'Reilly, 1990, X Toolkit Intrinsics Programming Manual, Volume Four, O'Reilly & Associates, Inc.
2. David Flanagan, 1991, Programmer's Supplement for Release 5, O'Reilly & Associates, Inc.
3. ทวีศักดิ์ กออนันตกูล และคณะ, 1991, คอมพิวเตอร์กับภาษาไทย: การพัฒนามาตรฐานเบื้องต้นสำหรับเทคโนโลยีสารสนเทศของไทย, ศูนย์เทคโนโลยีอิเลคทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม
4. พระยาอุปกิตศิลปสาร, 1990, หลักภาษาไทย, ไทยวัฒนาพานิช



ผู้ร่วมโครงการ:

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

ศูนย์วิจัยและพัฒนาปัญญาประดิษฐ์
สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี

ภาควิชาอิเล็กทรอนิกส์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง