# สัมมนา

# เรื่อง

# ระบบแปลภาษาด้วยวิธีสร้างและซ่อมแซม

**Generate and Repair Machine Translation**

**(Kanlaya Naruedomkul and Nick Cercone)**

**กิตติพจน์ เตชะเกิดกมล**

**สัมมนานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต**

**ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์**

**คณะวิทยาศาสตร์**

**สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง**

**ปีการศึกษา 2550**

# คำนำ

สัมมนาฉบับนี้ได้ถูกจัดทำขึ้นเพื่อแสดงถึงเทคโนโลยีในการแปลภาษาด้วยเครื่องจักรซึ่งในที่นี้ก็คือซอฟต์แวร์ เราจะเห็นได้ว่าปัจจุบันมีเวบไซต์และโปรแกรมต่างๆ เป็นจำนวนมากที่ใช้ระบบแปลภาษา กลุ่มผู้พัฒนาทั้งหลายต่างก็ใช้หลักการนี้ในการพัฒนาและออกแบบระบบแปลภาษาดังกล่าว จึงทำให้ผู้จัดทำสนใจและอยากศึกษาเทคโนโลยีเหล่านี้ เพื่อเป็นประโยชน์ในการศึกษาต่อภายภาคหน้า ซึ่งสัมมนาฉบับนี้อ้างอิงมาจากนิตยสาร “Computational Intelligence” ฉบับที่ 18 เล่มที่ 3 ปี 2002 (อ้างอิงจาก http://www.blackwellpublishing.com ) แม้จะนานไปบ้างแต่ก็เป็นทฤษฎีที่น่าสนใจชิ้นหนึ่งที่สมควรศึกษาเพื่อเป็นแบบอย่างต่อไป

# สารบาญ


| | |
|---|---|
| บทนำ (Introduction) | 1 |
| วิเคราะห์และแปลภาษา (Analysis Lite Machine Translation) | 2 |
| ปรับปรุงแก้ไข TC (Translation Candidate Evaluation) | 7 |
| วนซ้ำแก้ไข (Repair and Iterate) | 13 |
| บทสรุป (Concluding Remarks) | 16 |
| เอกสารอ้างอิง (References) | 17 |

# Generate and Repair Machine Translation
# ระบบแปลภาษาด้วยวิธีสร้างและซ่อมแซม

Kanlaya Naruedomkul
Department of Mathematics, Mahidol University, Bangkok, Thailand
Nick Cercone
Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada

## 1.บทนำ (Introduction)

Generate and Repair Machine Translation (GRMT) ถูกสร้างขึ้นโดยมีวัตถุประสงค์เพื่อเพิ่มความแม่นยำและความถูกต้องในการแปลภาษา (Machine Translation หรือ MT) ซึ่ง GRMT มีประสิทธิภาพอย่างยิ่งในการจัดการกับภาษาต่างๆ ได้หลายภาษาโดยไม่ขึ้นตรงต่อระบบระเบียบของภาษานั้นๆ ทำให้ GRMT เป็นหนึ่งในระบบที่ดีที่สุดในการจัดการแปลภาษา ตัวระบบถูกออกแบบมาให้ใช้งานและทำความเข้าใจได้ง่ายแต่ยังคงไว้ซึ่งระเบียบกฎเกณฑ์ของภาษาไม่เหมือนกับวิธีการอื่นๆ ที่กำจัดกฎเกณฑ์ทางภาษาเหล่านี้ทิ้งไป นอกจากนี้ GRMT ยังแยกการประมวลผลระหว่างภาษาตั้งต้น (Source Language หรือ SL) กับภาษาที่ต้องการ (Target Language หรือ TL) อย่างเด่นชัด เพื่อประโยชน์ในการประมวลผลพร้อมกันหลายๆ ภาษาอีกด้วย

GRMT จะทำการสร้างตัวแทนการแปลภาษา (Translation Candidate หรือ TC) ขึ้นมาแทนที่ตัวภาษาตั้งต้นเพื่อป้องกันการสูญเสียข้อมูลระหว่างการทำการแปลภาษาและเพื่อให้สะดวกต่อการจัดการหลายๆ ภาษา

ตัวแทนการแปลภาษา นั้นจะถูกสร้างขึ้นโดยดูจากความแตกต่างระหว่างภาษาเป็นหลักโดยไม่มีการวิเคราะห์ปัญหาที่ยุ่งยากอะไรทั้งสิ้น ทำให้การทำงานของระบบเป็นไปอย่างราบรื่นและรวดเร็ว หลังจากที่ตัวแทนการแปลภาษาถูกสร้างขึ้นแล้ว GRMT จะทำการตรวจสอบความถูกต้องของตัวแทนการแปลภาษาอีกครั้ง ซึ่งอาจส่งผลให้มีการแก้ไขปรับปรุงตัวแทนการแปลภาษาอีกครั้งหากจำเป็น หลังจากนั้นระบบจะทำการวิเคราะห์และซ่อมแซม TC แบบนี้ไปเรื่อยๆ จนกระทั่งตัวแทนการแปลภาษามีความหมายเหมือนกับภาษาตั้งต้น

GRMT ประกอบไปด้วยขั้นตอนทั้งหมด 3 ขั้นตอนด้วยกันคือ วิเคราะห์และปรับปรุงภาษา (Analysis Lite Machine Translation หรือ ALMT) ปรับปรุงแก้ไขตัวแทนการแปลภาษา (Translation Candidate Evaluation หรือ TCE) และวนซ้ำแก้ไข (Repair and Iterate หรือ RI)

GRMT มีวิธีการจัดการกับข้อแตกต่างของแต่ละภาษาในวิธีที่เป็นเอกเทศ โดยการจัดภาษาต่างๆ ให้เป็นหมวดหมู่ซึ่งพิจารณาจาก รูปของกริยา(tense) กรรมวาจก(passive voice) เป็นต้น ตัวอย่างหมวดหมู่ของภาษาเหล่านี้ได้แก่ ภาษาอังกฤษ ภาษาฝรั่งเศส ภาษาสเปน ถูกจัดให้อยู่ใน

หมวดหมู่เดียวกัน หรืออีกหมวดหมู่คือ ภาษาไทย ภาษาญี่ปุ่น ภาษาจีน เป็นต้น การที่จะแปลกลับไปกลับมาระหว่างภาษาใน 2 กลุ่มนี้ หากเราใช้วิธีการแปลงโดยตรงจะต้องใช้ ตัววิเคราะห์ภาษาตั้งต้นทั้งหมด 6 ตัว ตัววิเคราะห์ภาษาที่ต้องการอีก 6 ตัว และมีกฎในการแปลงทั้งหมด 18 ข้อ แต่หากเราใช้ GRMT มาจัดการแล้ว เราใช้เพียงแค่ตัววิเคราะห์ภาษา 6 ตัว และตัวจัดการข้อจำกัดของภาษาอีก 2 ตัวเท่านั้น

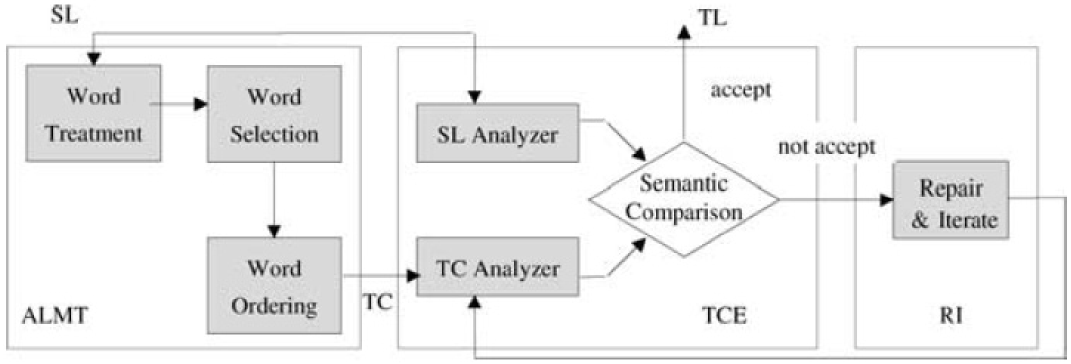


FIGURE 1. GRMT Architecture.

**2. วิเคราะห์และแปลภาษา (Analysis Lite Machine Translation หรือ ALMT)**

การที่ ALMT จะสร้างตัวแทนการแปลภาษาขึ้นมานั้น ประกอบไปด้วยขั้นตอนย่อยทั้งหมด 3 ขั้นตอน คือ การจัดการคำเบื้องต้น (Word Treatment) การเลือกคำ (Word Selection) และการเรียงลำดับคำ (Word Ordering)

**การจัดการคำเบื้องต้น (Word Treatment)**

ในขั้นตอนของการจัดการคำเบื้องต้นประกอบไปด้วยขั้นตอนย่อยอีก 2 ขั้นตอนคือ การจัดการกับข้อจำกัดของภาษา เป็นการย่นย่อขนาดของข้อความซึ่งจะคัดเลือกคำจากภาษาที่ต้องการที่สอดคล้องกับภาษาตั้งต้น ส่วนอีกขั้นตอนหนึ่งคือ การสืบค้นข้อมูลพจนานุกรมซึ่งเป็นการนำเอาคำแต่ละคำในภาษาที่ต้องการมากำหนดลงไปให้สอดคล้องกับภาษาตั้งต้น

**การจัดการกับข้อจำกัดของภาษา**

ข้อจำกัดของภาษาตั้งต้นแต่ละภาษาคือลักษณะพิเศษที่มีอยู่ในภาษานั้นๆ เช่น ความเป็นพหูพจน์ (Plurality) รูปกาลของประโยคที่เกิดอย่างต่อเนื่อง (Continuous Tense) ฯลฯ

จากการที่คำในภาษาไทยนั้นไม่มีการผันตามรูปกาลหรือความเป็นพหูพจน์ของคำ จึงมีการนำเอาส่วนประกอบของความเป็นพหูพจน์และรูปกาลออกไปในขั้นตอนนี้ หลังจากที่นำส่วนประกอบเหล่านี้ออกไปจากคำแล้ว จึงมีการนำส่วนประกอบเหล่านี้มาเพิ่มเข้าไปในประโยคอีกทีเพื่อคงสภาพไว้ซึ่งความเดิมของประโยค ดังจะเห็นได้จาก ตารางที่ 1 ด้านล่าง

TABLE 1. Some SL Constraints considered in the English–Thai MT system

| SL Constraints | Descriptions | Examples |
|---|---|---|
| Plurality | noun_(e)s → noun + plural | The books . . . → The book + plural |
| Continuous tense | V to be + V_ing → V + ing | I am swimming. → I + ing + swim |
| Passive voice | be + V3 → passive + V | He was arrested. → He passive + arrest |
| Negative | do + not → not | He does not eat → He not eat |

จากตารางที่ 1 จะเห็นได้ว่า ในการจัดการรูปพหูพจน์ ระบบจะทำการตัดอักษร s ซึ่งแสดงความเป็นพหูพจน์ทิ้งไป และเพิ่มส่วนของ plural ลงไปเพื่อคงสภาพไว้ซึ่งความเดิมของประโยค ส่วนรูปแบบอื่นๆ ก็ทำเช่นเดียวกัน

**การสืบค้นข้อมูลพจนานุกรม**

หลังจากที่เราทำการจัดการกับข้อจำกัดของภาษาแล้ว คำในภาษาตั้งต้นทุกคำจะถูกนำมาใช้เป็นคำสำคัญในการค้นหาคำที่สอดคล้องกันกับภาษาที่ต้องการ ถ้าคำสำคัญดังกล่าวปรากฏในฐานข้อมูลพจนานุกรม คำที่เป็นไปได้ทั้งหมดที่พบจะถูกนำมาเชื่อมต่อกับคำในภาษาตั้งต้นนั้นๆ แต่ถ้าไม่พบคำดังกล่าวในฐานข้อมูลพจนานุกรม ระบบจะทำการจัดการกับข้อจำกัดของภาษาและค้นหาในฐานข้อมูลพจนานุกรมอีกครั้ง ส่งผลให้ส่วนที่เป็นการผันรูปของคำก็ถูกตัดทอนออกไปทำให้การค้นหาข้อมูลครั้งต่อไปเร็วขึ้นอีกด้วย ผลลัพธ์ที่ได้จากขั้นตอนนี้คือคำทั้งหมดที่สอดคล้องกับภาษาตั้งต้นเป็นภาษาที่ต้องการตามด้วยตัวเลขที่ใช้ในการแสดงความสัมพันธ์ของกลุ่มคำ ซึ่งต่อไปนี้เราจะเรียกตัวเลขเหล่านี้ว่า WordAsso (Word Association Number)

**การเลือกคำ (Word Selection)**

การคัดเลือกคำนั้นจะอาศัยความสัมพันธ์ของคำในเชิงความหมายเป็นหลัก ซึ่งในที่นี้เราจะใช้ WordAssoซึ่งเป็นค่าตัวเลขแบบพิเศษในการวิเคราะห์

WordAsso คือค่าตัวเลขประเภทหนึ่งที่ถูกกำหนดให้กับหมวดหมู่ของคำ ซึ่งมีลักษณะดังตัวอย่างต่อไปนี้

เช่น คำที่มีหมายเลข WordAsso เป็น 1-1-1-1 หมายถึงอยู่ในหมวดหมู่ของมนุษย์

คำที่มีหมายเลข WordAsso เป็น 1-1-1-1-2 แสดงว่าคำนี้อยู่ลึกลงไปภายใต้หมวดหมู่ของมนุษย์ ซึ่งในที่นี้หมายถึงหมวดหมู่ "ผู้หญิง"

เป็นต้น

Example 1:
SL: Algebraic symbols are used when you do not know what you are talking about.
SL Cons- output: Algebraic plural symbol passive use when you not know what you ing talk about.

FIGURE 2. SL constraints applications examples.

**การเรียงลำดับคำ (Word Ordering)**

ในขั้นตอนการเรียงลำดับคำนี้ประกอบไปด้วย 2 ขั้นตอนย่อยด้วยกันได้แก่ การเพิ่มคำ (Word Addition) และการเรียงคำ (Word Ordering) โดยรวมแล้วก็คือการจัดการกับไวยากรณ์ที่แตกต่างกันระหว่างภาษาตั้งต้นกับภาษาที่ต้องการเพื่อสร้างตัวแทนการแปลภาษาที่สมบูรณ์ คำบางคำที่ช่วยดำรงไว้ซึ่งความหมายเดิมของภาษาตั้งต้นและมีประโยชน์ช่วยในการทำให้รูปไวยากรณ์ถูกต้องจะถูกเพิ่มเข้ามา ในกรณีของภาษาที่ไม่มีการผันรูปของคำ เช่น ภาษาไทย ภาษาจีน ภาษาญี่ปุ่น เราจะเพิ่มคำประเภทลักษณนามเข้าเพื่อทำให้เป็นรูปอดีตกาลหรือพหูพจน์

เมื่อเพิ่มคำที่จำเป็นเข้าไปจนครบแล้ว ALMT จะทำการจัดเรียงคำที่จำเป็นตามหลักโครงสร้างไวยกรณ์โดยไม่มีการวิเคราะห์ใดๆ ทั้งสิ้น ในด้านโครงสร้างของภาษาไทยกับภาษาอังกฤษนั้นมีสิ่งที่เหมือนกันคือ รูปประโยคจะประกอบไปด้วย ประธาน กริยา และกรรม ตามลำดับ แต่เมื่อพิจารณาให้ละเอียดลงไปแล้วจะเห็นว่ามีส่วนที่แตกต่างกันอยู่มาก เช่น ในภาษาไทย ประธานจะต้องตามด้วยคำคุณลักษณะ คำแสดงความเป็นเจ้าของ และคำบ่งชี้ ตามลำดับ ซึ่งคำว่า can , could และ must นั้นถูกจัดเรียงในเชิงตรงข้ามกัน

Example 2: An old woman lived in the cottage, with a fat black cat and a plump brown hen.

TC: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สีดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง[3]
CT: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สีดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง
(phûujiˇη[4]-woman) (kε`ε-old) (khon-clas) (ny`η-an) (dâj-past) (ju`u-live) (naj-in) (krathə̂m -cottage) (laˇη-clas) (nán-the) (kàb-with) (mεεw-cat) (siˇidam-black) (?ûan-fat) (tua-clas) (ny`η-a) (lε`-and) (kàj-hen) (siˇinámtaan-brown) (?ùab-plump) (tua-clas) (ny`η-a)

TABLE 2. ALMT Steps Applied to the Sentence of Example 2.

| English | Constraint Application Output | Word Treatment Output | Selected Word in Thai | Word Addition | Word Ordering |
|---|---|---|---|---|---|
| An | An | หนึ่ง (nyˋŋ) | หนึ่ง (nyˋŋ) | | ผู้หญิง (phûujiˇŋ) |
| | | | | คน (khon) | แก่ (kɛˋɛ) |
| old | old | แก่ (kɛˋɛ), เก่า (kàw) | แก่ (kɛˋɛ), | | คน (khon) |
| woman | woman | ผู้หญิง (phûujiˇŋ) | ผู้หญิง (phûujiˇŋ) | | หนึ่ง (nyˋŋ) |
| | | | | ได้ (dâj) | ได้ (dâj) |
| lived | live | อยู่ (juˋu), มีชีวิต (miichiiwíd), ดำรงชีวิต (damroŋchiiwíd) | อยู่ (juˋu) | | อยู่ (juˋu) |
| in | in | ใน (naj), เข้ามา(khâwmaa), เข้าไป(khâwpaj), อยาง(jàaŋ) | ใน (naj) | | ใน (naj) |
| the | the | นั้น (nán) | นั้น (nán) | | กระท่อม (krathəˆm) |
| | | | | หลัง (laˇŋ) | หลัง (laˇŋ) |
| cottage | cottage | กระท่อม (krathəˆm) | กระท่อม (krathəˆm) | | นั้น (nán) |
| with | with | กับ (kàb), ด้วย (dûaj), ซึ่งมี (syˆŋ) | กับ (kàb) | | กับ (kàb) |
| a | a | หนึ่ง (nyˋŋ) | หนึ่ง (nyˋŋ) | | แมว (mɛɛw) |
| | | | | ตัว(tua) | สีดำ (siˇidam) |
| fat | fat | อ้วน (?ûan), ไขมัน (khaˇjman) | อ้วน (?ûan) | | อ้วน (?ûan) |
| black | black | ดำ (dam), สีดำ (siˇidam) | สีดำ (siˇidam) | | ตัว (tua) |
| cat | cat | แมว (mɛɛw), ผู้หญิงไม่ดี (phûujiˇŋmâjdìi) | แมว (mɛɛw) | | หนึ่ง (nyˋŋ) |
| and | and | และ (lɛˋ) | และ (lɛˋ) | | และ (lɛˋ) |
| a | a | หนึ่ง (nyˋŋ) | หนึ่ง (nyˋŋ) | | ไก่ (kàj) |
| | | | | ตัว (tua) | สีน้ำตาล (siˇinámtaan) |
| plumb | plumb | อวบ (?ùab) | อวบ (?ùab) | | อวบ (?ùab) |
| brown | brown | สีน้ำตาล(siˇinámtaan), เกรียม(kriam), คล้ำ (khlám) | สีน้ำตาล(siˇinámtaan) | | ตัว (tua) |
| hen | hen | ไก่ (kàj) | ไก่ (kàj) | | หนึ่ง (nyˋŋ) |

ตารางที่ 2 เป็นผลลัพธ์จากการทำงานของ ALMT ในตัวอย่างที่ 2

- คอลัมน์แรกจะแสดงให้เห็นถึงภาษาตั้งต้นที่ยังไม่ได้ผ่านการประมวลผลใดๆ ทั้งสิ้น

- คอลัมน์ที่สองเป็นภาษาตั้งต้นที่ผ่านการจัดการคำเบื้องต้นแล้ว ซึ่งคำทั้งหมดในคอลัมน์ที่สองนี้จะถูกนำไปใช้เป็นคำสำคัญในการค้นหาจากฐานข้อมูลพจนานุกรมเพื่อหาคำในภาษาไทย
- คอลัมน์ที่สามแสดงถึงคำทั้งหมดที่ความหมายเหมือนหรือใกล้เคียง
- คอลัมน์ที่สี่แสดงถึงคำที่เลือกใช้ การจะเลือกคำที่เหมาะสมสำหรับคำว่า old และ in นั้น จะพิจารณาจากความสัมพันธ์ระหว่างความหมายของคำแต่ละคำ แต่เราไม่สามารถใช้วิธีการเช่นนี้กับคำว่า live และ with ได้เพราะคำเหล่านี้ไม่มีความสัมพันธ์กับคำอื่นๆ (ดูตาม WordAsso) ดังนั้นเราจึงใช้วิธีเลือกคำแรกสุดที่ปรากฏแทน
- คอลัมน์ที่ห้าแสดงถึงคำพิเศษที่จำเป็นต้องเพิ่มเข้ามา
    - คำว่า “ได้” ถูกเพิ่มเข้ามา เพื่อแสดงรูปอดีต
    - คำว่า “คน” “หลัง” “ตัว” ถูกเพิ่มเข้ามาเพื่อให้เป็นไปตามหลักโครงสร้างไวยากรณ์ของภาษาไทยที่ต้องการลักษณนาม
    - คำบ่งชี้ในที่นี้คือ a กับ an ซึ่งมีความหมายว่า “หนึ่ง” แต่ในหลักไวยากรณ์ภาษาไทยเราจะต้องมีลักษณนามตามหลังด้วยซึ่งเราจะดูตาม WordAsso ดังนี้คือ

1.คำว่า “ผู้หญิง” (1-1-1-1-1-2) เป็นหมวดหมู่ย่อยของ “มนุษย์” (1-1-1-1) ดังนั้นคำนามที่อยู่ในหมวดหมู่ 1-1-1-1 หรือคำว่า “คน” จึงเหมาะสมที่จะนำมาใช้

```
clf_rel('1-1-1-1',[[wasso('คน',['2-4-2-1-1-1'])]]).
clf_rel('1-1-1-2',[[wasso('ตัว',['2-4-2-1-1-2'])]]).
clf_rel('1-1-2-1-6-1',[[wasso('แห่ง',['2-4-2-1-2-2'])]]).
clf_rel('2-4-3-1-2',[[wasso('วัน',['2-4-2-3'])]]).
clf_rel('1-1-2-1-1-2-2',[[wasso('หลัง',['2-4-2-1-2-10'])]]).
```

FIGURE 3. Examples of classifier relations.

จากแผนภาพคำว่า “คน” ที่มี WordAsso หมายเลข 2-4-2-1-1-1 มีความสัมพันธ์กับ WordAsso 1-1-1-1 (มนุษย์) ดังรูปข้างต้น

2.คำว่า “แมว” และคำว่า “ไก่” ถูกจัดอยู่ในหมวดหมู่ของ “สัตว์เลี้ยงลูกด้วยนม” (1-1-1-2-1-1) กับหมวดหมู่ของ “สัตว์ปีก” (1-1-1-2-1-2-2) ซึ่งต่างก็อยู่ภายใต้หมวดหมู่ของ “สัตว์” (1-1-1-2) จึงเลือกใช้คำว่า “ตัว” มาเป็นลักษณนาม

- “the” เป็นคำบ่งชี้ที่มีความสัมพันธ์กับกระท่อมซึ่งมีความหมายในภาษาไทยว่า “นั้น” มีความต้องการลักษณนามด้วยเช่นกัน ในที่นี้จากความสัมพันธ์ของคำว่า “the” กับคำว่า “กระท่อม” (1-1-2-1-1-2-2) เราจึงใช้คำว่า “หลัง” เป็นลักษณนาม

**3. ปรับปรุงแก้ไข TC (Translation Candidate Evaluation หรือ TCE)**

หน้าที่หลักๆ ของ TCE นั้นจะเป็นการตรวจสอบว่าตัวแทนการแปลภาษาที่สร้างขึ้นมานั้นยังคงมีความหมายเดียวกับภาษาตั้งต้นอยู่ ซึ่งในขั้นตอนนี้จะทำการตรวจสอบความหมายของทั้งภาษาตั้งต้นและตัวแทนการแปลภาษาไปพร้อมๆ กันทีละส่วน หากความหมายที่ได้ของทั้งสองมีค่าออกมาเหมือนกัน แสดงว่าการแปลภาษาอาจประสบผลสำเร็จ แต่ทว่าหากผลลัพธ์ทั้งสองออกมาไม่เหมือนกัน จะมีการซ่อมแซม TC ในขั้นตอนที่สามคือ ขั้นวนซ้ำแก้ไข (Repair and Iterate)

ในการทำ TCE ประกอบไปด้วยขั้นตอนทั้งหมด 2 ขั้นตอนด้วยกัน คือ ขั้นตอนการตรวจสอบ (The Analyzer) และขั้นตอนการเปรียบเทียบความหมาย (Semantic Comparison)

**ขั้นตอนการตรวจสอบ (The Analyzer)**

ในขั้นตอนนี้ระบบจะทำการตรวจสอบโครงสร้างไวยากรณ์และความหมายของ TC ว่าถูกต้องตามโครงสร้างไวยากรณ์หรือไม่ และยังคงไว้ซึ่งความหมายเดิมของภาษาตั้งต้นหรือไม่ ซึ่งในขั้นตอนนี้แบ่งย่อยลงไปได้อีก 2 ขั้นตอนคือ การวิเคราะห์คำ (Parsing) และการดึงความหมาย (Semantic Extraction)

ในขั้นตอนการวิเคราะห์คำ เราจะทำการวิเคราะห์ไปพร้อมๆ กันทั้ง SL และ TC ซึ่งไวยากรณ์ที่เราจะใช้ในการวิเคราะห์นี้เป็นไวยากรณ์ของภาษาอังกฤษและภาษาไทยที่ได้จากวิธีที่เรียกว่า Head-Driven Phrase Structure Grammar (HPSG) ข้อดีของ HPSG คือ มีข้อมูลของแต่ละคำให้ใช้สอยเป็นจำนวนมาก มีกฎของไวยากรณ์น้อย และมีการใช้โครงสร้างบางส่วนร่วมกัน

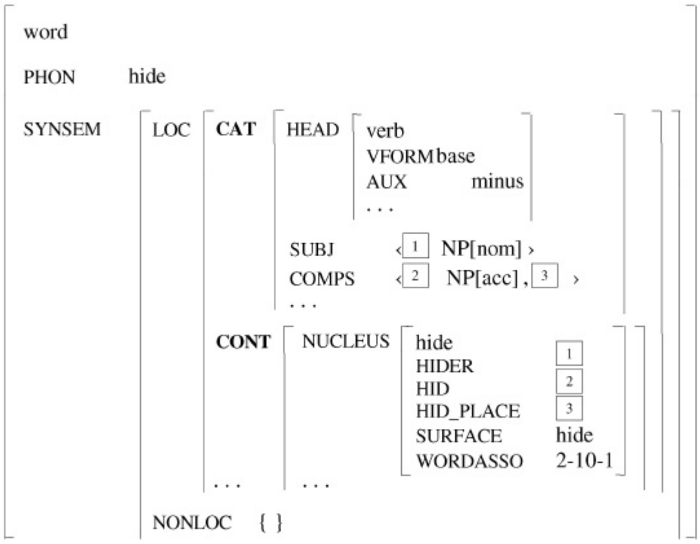


FIGURE 4. The lexical entry for *hide*.

HPSG จะมองว่าคำแต่ละคำประกอบไปด้วยข้อมูลโครงสร้างที่ซับซ้อนแต่มีข้อมูลเฉพาะบางส่วนที่เหมือนกันซึ่งภายในข้อมูลของแต่ละคำที่ถูกสร้างขึ้นด้วย HPSG จะแบ่งออกแบบหมวดหมู่ย่อยๆ ลงไป ดังแผนภาพด้านบนมีลักษณะหนึ่งที่เรียกว่า ประเภท หรือ CATEGORY(CAT) และเนื้อหา หรือ CONTENT(CONT) ซึ่งแสดงให้เห็นถึงโครงสร้างและความหมายของคำนั้นๆ รูปข้างต้น (รูปที่ 4) เป็นแผนภาพที่แสดงให้เห็นถึงส่วนหนึ่งของข้อมูลในคำว่า “hide” ซึ่งสามารถอธิบายข้อมูลได้ดังนี้

1. “hide” มีสภาพเป็นคำหนึ่งคำ (word)
2. “hide” เป็น verb ที่ยังไม่มีการผันรูป (verb, VFORM:base)
3. “hide” ไม่ได้เป็นกริยานุเคราะห์หรือกริยาช่วย (AUX:minus)
4. “hide” มีประธานเป็นนามหรือนามวลี (SUBJ:NP[nom]) และประธานตัวนี้จะมีความสัมพันธ์กับ CONT ที่มีค่า hide คือเป็น “HIDER”
5. ส่วนเติมเต็มคำแรกของ hide คือ คำนามที่เป็นกรรม (COMPS:NP[acc]) ซึ่งจะต้องรวมเข้าด้วยกันกับ CONT ที่มีค่า hide ซึ่งก็คือ HID
6. ส่วนเติมเต็มคำที่สองจะต้องรวมเข้าด้วยกันกับ CONT ที่มีค่า hide ซึ่งก็คือ HID_PLACE

ภาพที่ 5 แสดงให้เห็นถึงการวิเคราะห์คำในตัวอย่างที่ 3 และตัวแทนการแปลภาษา

ในขั้นตอนการดึงความหมาย เราจะทำการดึงเอาข้อมูลที่เป็นความหมายของทั้งภาษาตั้งต้นและภาษาที่ต้องการมาเปรียบเทียบซึ่งกันและกัน

CONT CONX และ QSTORE เป็นตัวที่ใช้เก็บความหมายของคำ จากภาพที่ 4 เราจะสามารถดึงความหมายออกมาได้ดังนี้

- คำว่า “hide” (ซ่อน) สอดคล้องกับบริเวณ CONT (กล่องหมายเลข 1) นั่นก็คือ (ผู้ซ่อน) ซ่อน (hid บริเวณกล่องหมายเลข 2 ) ใน HID_PLACE (ที่ซ่อน ในบริเวณกล่องหมายเลข 3)
- SURFACE แสดงให้เห็นถึงรูปฟอร์มของคำ ซึ่ง SURFACE และ WORDASSO ในที่นี้จะใช้ในขั้นตอนของการวนซ้ำแก้ไข
- QSTORE เป็นส่วนที่ใช้เก็บตัวบ่งปริมาณ
- CONX แสดงถึงข้อมูลทางภาษา

ตัวอย่างของ QSTORE และ CONX สามารถดูได้ที่ภาพที่ 11

Example 3: The ugly duckling hides his head under his wing.

SL: The ugly duckling hides his head under his wing.
TC: ลูกเป็ด ขี้เหร่ ตัว นั้น ซ่อน หัว ของเขา ใต้ ปีก ของเขา
(lûugpèd-duckling) (khîirèe-ugly) (tua-clas) (nán-the) (səˆən-hide) (huˇa-head)
(khəˇəŋkhaw-his) (tâaj-under) (pìig-wing) (khəˇəŋkhaw-his)

```
STRING:
0 the 1 ugly 2 duckling 3 hides 4 his 5 head
6 under 7 his 8 wing 9
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                    NUCLEUS beauty
                        WORDASSO 13-8-4
                        INSTANCE [4]
                        SURFACE ugly
. . . .
SYNSEM synsem
   LOC loc
      CAT cat
         COMPS e_list
         HEAD verb
            AGR agr
               GENN gend
               NUM num
               PER per
            AUX minus
            INV minus
            LANG eng
            MOD none
            NEG boolean
            PRD boolean
            VFORM fin
         MARKING unmarked
         SPR list_synsem
         SUBJ e_list
      CONT psoa
         NUCLEUS hide2
               WORDASSO 2-10-1
               HID1 [1]
               HID_PLACE [3]
               HIDER1 [4]
               SURFACE hide
         QUANTS e_list
      CONX conx
. . .
```

```
STRING:
0 ลูกเป็ด 1 ขี้เหร่ 2 ตัว 3 นั้น 4 ซ่อน 5 หัว
6 ของเขา 7 ใต้ 8 ปีก 9 ของเขา 10
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                    NUCLEUS beauty
                        WORDASSO 13-8-4
                        INSTANCE [4]
                        SURFACE ขี้เหร่
. . .
SYNSEM synsem
   LOC loc
      CAT cat
         COMPS e_list
         HEAD verb
            AGR agr
               GENN gend
               NUM num
               PER per
            AUX minus
            INV minus
            LANG thai
            MOD none
            NEG boolean
            PRD boolean
            VFORM bse
         MARKING unmarked
         SPR ne_list_synsem
            HD synsem
               LOC loc
                  CAT cat
                     COMPS list_synsem
                     . . .
                     MARKING comp
                     SPR list_synsem
                     SUBJ list_synsem
. . .
         SUBJ e_list
      CONT psoa
         NUCLEUS hide2
               WORDASSO 2-10-1
               HID1 [1]
               HID_PLACE [3]
               HIDER1 [4]
               SURFACE ซ่อน
         QUANTS e_list
      CONX conx
. . .
```

FIGURE 5. The syntax of Example 3 and that of its translation.

**การเปรียบเทียบความหมาย (Semantic Comparison)**

ในขั้นตอนนี้เราจะพิจารณาจาก CONT CONX และ QSTORE ถ้าความหมายที่วิเคราะห์ได้จาก TC ในส่วนนี้ยังคงมีความหมายเดียวกันกับภาษาตั้งต้นจะไม่มีการซ่อมแซมตัวแทนการแปลภาษาแต่อย่างใด แต่ถ้ามีบางส่วนที่แตกต่างกัน TCE จะทำการสร้างข้อมูลชนิดหนึ่งที่แตกต่างกับตัวแทนการแปลภาษา ซึ่งจะใช้ต่อไปในขั้นตอนวนซ้ำแก้ไข จากข้อมูลในรูปที่ 6-7 (เป็นข้อมูลที่ได้จากตัวอย่างที่ 3) TCE จะเริ่มเปรียบเทียบความหมายจาก CONT ซึ่งจะเห็นว่าในส่วนนี้จะแสดงให้เห็นถึงความสัมพันธ์ของ hide2 กับสิ่งที่มีความสัมพันธ์กัน คือ HIDER HID และ HID_PLACE เมื่อค่าในส่วนของ CONT ของทั้งภาษาตั้งต้นและภาษาที่ต้องการเหมือนกันแล้ว TCE จะเริ่มทำการตรวจสอบค่า QSTORE และ CONX ต่อ เพื่อให้รู้แน่ชัดว่ามีความสัมพันธ์กับคำเดียวกันหรือเปล่า

- กล่องหมายเลข 1 ใน CONT แสดงให้เห็นถึงการเชื่อมต่อระหว่าง bodypart ในหมวด 1-4-1-1-1 ซึ่งก็คือคำว่า “head” ในภาษาตั้งต้น และเป็นคำว่า “หัว” ในภาษาที่ต้องการ
- กล่องหมายเลข 2 แสดงให้เห็นถึงการเชื่อมต่อระหว่าง bodypart ในหมวด 1-4-1-1-3-1 ซึ่งก็คือ “wing” ในภาษาตั้งต้น และเป็นคำว่า “ปีก” ในภาษาที่ต้องการ
- กล่องข้อความที่ 3 แสดงให้เห็นว่าเป็นสัตว์ (animal) ในหมวด 1-1-1-2-1-2-1 ซึ่งก็คือ “duckling” ในภาษาตั้งต้นและเป็นคำว่า “เป็ด” ในภาษาที่ต้องการ

CONX ในตัวอย่างนี้ไม่มีข้อมูล ซึ่งเราจะเห็นได้ว่าภาษาตั้งต้นและภาษาที่ต้องการในที่นี้ไม่มีข้อแตกต่าง จึงสรุปว่าตัวแปรการแทนภาษาที่สร้างขึ้นมานี้เหมาะสมแล้ว

ในกรณีของตัวอย่างที่ 4 (ภาพที่ 8) เราจะเห็นได้ว่า ค่า WordAsso ของภาษาตั้งต้นกับภาษาที่ต้องการมีค่าต่างกัน ใน SL มีค่าคือ 2-1-19 ซึ่งเป็นคำว่า “cover” แต่ใน TC มีค่าคือ 2-1-19-2 ซึ่งเป็นคำว่า “คลุม” เราจะสังเกตว่าระหว่าง 2 คำนี้มีค่าตัวเลข WordAsso ต่างกัน ถึงแม้คำทั้งสองจะมีความหมายเหมือนกันแต่ว่ามีการใช้งานจริงต่างกัน เช่น ประโยคที่ว่า John covers his ears with his hands ซึ่งคำว่า cover ในประโยคนี้แปลได้ว่า “ปิด” ในภาษาไทย แต่เราจะถือว่าความหมายของคำไม่แตกต่างกันหากมันจัดอยู่ในเลขหมวดหมู่หลักเดียวกัน

```
STRING:
0 the 1 ugly 2 duckling 3 hides 4 his 5 head 6 under 7 his 8 wing 9
CATEGORY:

phrase
QSTORE ne_set_quant
  ELT quant
    DET the
    RESTIND nom_obj
      INDEX [0] ref
        IGENN gend
        INUM num
        IPER per
      MODE mode
      RESTR ne_set_psoa
        ELT psoa
          NUCLEUS possess
            WOROASSO wordasso
            POSSESSED [1] ref
              IGENN gend
              INUM sg
              IPER per
            POSSESSOR [0]
            SURFACE surface
          QUANTS e_list
        ELTS ne_set_psoa
          ELT psoa
            NUCLEUS bodypart
              WORDASSO 1-4-1-1-1
              INSTANCE [1]
              SURFACE head
                AGR agr
                  GENN gend
                  NUM num
                  PER per
                LANG lang
            QUANTS e_list
          ELTS e_set
  ELTS ne_set_quant
    ELT quant
      DET the
      RESTIND nom_obj
        INDEX [2] ref
          IGENN gend
          INUM num
          IPER per
        MODE mode
        RESTR ne_set_psoa
          ELT psoa
            NUCLEUS possess
              WOROASSO wordasso
              POSSESSED [3] ref
                IGENN gend
                INUM sg
                IPER per
                                    . . .
              POSSESSOR [2]
              SURFACE surface
            QUANTS e_list
          ELTS ne_set_psoa
            ELT psoa
              NUCLEUS bodypart
                WORDASSO 1-4-1-1-3-1
                INSTANCE [3]
                SURFACE wing
              QUANTS e_list
            ELTS e_set
    ELTS ne_set_quant
      ELT quant
        DET the
        RESTIND nom_obj
          INDEX [4] ref
            IGENN gend
            INUM sg
            IPER per
          MODE reff
          RESTR ne_set_psoa
            ELT psoa
              NUCLEUS beauty
                WORDASSO 13-8-4
                INSTANCE [4]
                SURFACE ugly
              QUANTS e_list
            ELTS ne_set_psoa
              ELT psoa
                NUCLEUS animal
                  WORDASSO 1-1-1-2-1-2-1
                  INSTANCE [4]
                  SURFACE duckling
                QUANTS e_list
              ELTS e_set
      ELTS e_set
. . .
CONT psoa
  NUCLEUS hide2
    WORDASSO 2-10-1
    HID1 [1]
    HID_PLACE [3]
    HIDER1 [4]
    SURFACE hide
  QUANTS e_list
CONX conx
  BACKGR e_set
```

FIGURE 6. The semantic information of the SL parse in Example 3.

```
STRING:
0 ลูกเป็ด 1 ขี้เหร่ 2 ตัว 3 นั้น 4 ซ่อน 5 หัว 6 ของเขา 7 ใต้ 8 ปีก 9 ของเขา 10
CATEGORY:

phrase
QSTORE ne_set_quant
   ELT quant
     DET the
     RESTIND nom_obj
         INDEX [0] ref
            IGENN gend
            INUM num
            IPER per
         MODE mode
         RESTR ne_set_psoa
            ELT psoa
              NUCLEUS possess
                  WORDASSO wordasso
                  POSSESSED [1] ref
                       IGENN gend
                       INUM sg
                       IPER per
                  POSSESSOR [0]
                  SURFACE surface
              QUANTS e_list
            ELTS ne_set_psoa
              ELT psoa
                NUCLEUS bodypart
                    WORDASSO 1-4-1-1-1
                    INSTANCE [1]
                    SURFACE หัว
                QUANTS e_list
              ELTS e_set
   ELTS ne_set_quant
     ELT quant
       DET the
       RESTIND nom_obj
           INDEX [2] ref
              IGENN gend
              INUM num
              IPER per
           MODE mode
           RESTR ne_set_psoa
              ELT psoa
                NUCLEUS possess
                    WORDASSO wordasso
                    POSSESSED [3] ref
                         IGENN gend
                         INUM sg
                         IPER per
                    POSSESSOR [2]
                    SURFACE surface
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS bodypart
                      WORDASSO 1-4-1-1-3-1
                      INSTANCE [3]
                      SURFACE ปีก
                  QUANTS e_list
                ELTS e_set
     ELTS ne_set_quant
       ELT quant
         DET the
         RESTIND nom_obj
             INDEX [4] ref
                IGENN gend
                INUM sg
                IPER per
             MODE reff
             RESTR ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [4]
                      SURFACE ขี้เหร่
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [4]
                        SURFACE ลูกเป็ด
                    QUANTS e_list
                  ELTS e_set
       ELTS e_set
 . . .
CONT psoa
  NUCLEUS hide2
      WORDASSO 2-10-1
      HID1 [1]
      HID_PLACE [3]
      HIDER1 [4]
      SURFACE ซ่อน
  QUANTS e_list
CONX conx
  BACKGR e_set
 . . .
```

FIGURE 7. The semantic information of the TC parse in Example 3.

Example 4:

SL: John covers his head with his coat.
TC: จอห์น คลุม หัว ของเขา ด้วย เสื้อคลุม ของเขา
(cɔɔn-John) (khlum-cover) (huˇa-head) (khəˇəŋkhaw-his) (dûaj-with)
(syˆakhlum-coat)( khəˇəŋkhaw-his)

```
STRING:
0 john 1 covers 2 his 3 head 4 with 5 his 6 coat 7
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
        CONT psoa
           NUCLEUS cover2
                 WORDASSO 2-1-19
                 COVERED [1]
                 COVERED_PLACE [3]
                 COVERER1 [4] ref
                        IGENN masc
                        INUM sg
                        IPER per
                 SURFACE cover
           QUANTS e_list
        CONX conx
. . .
```

```
STRING:
0 จอห์น 1 คลุม 2 หัว 3 ของเขา 4 ด้วย 5 เสื้อคลุม 6 ของเขา 7
CATEGORY:

phrase
QSTORE ne_set_quant

. . .
        CONT psoa
           NUCLEUS cover2
                 WORDASSO 2-1-19-2
                 COVERED [1]
                 COVERED_PLACE [3]
                 COVERER1 [4] ref
                        IGENN masc
                        INUM sg
                        IPER per
                 SURFACE คลุม
           QUANTS e_list
        CONX conx
. . .
```

FIGURE 8. The CONT value of the SL parse and that of the TC of Example 4.

TABLE 3. Class of *cover*.

| WordAsso and Class description | Word |
|---|---|
| 2-1-19 Cover as to place something on or over to protect or hide | cover |
| 2-1-19-1 Cover as to cover eyes, ears | ปิด (pid) |
| 2-1-19-2 Envelop | ดคลุม (khlum) |

**4. วนซ้ำแก้ไข (Repair And Iteration หรือ RI)**

หน้าที่หลักของขั้นตอนนี้ก็คือ ซ่อมแซมประโยค และส่งตัวแทนการแปลภาษากลับไปยัง TCE (หากจำเป็น) หลังจากนั้น TCE จะทำการเปรียบเทียบความหมายของกับตัวแทนการแปลภาษาว่ายังต่างกันอยู่หรือไม่ ซึ่ง RI จะทำการตรวจสอบจากตัวแทนการแปลภาษาที่ได้จากขั้นตอน TCE ซึ่งอยู่ในรูปที่แตกต่างกันกับตัวแทนการแปลภาษาที่ใช้จริง ส่งผลให้ RI สามารถรู้ได้ว่ามีการแปลที่ผิดพลาดหรือไม่ หลังจากแก้ไขในส่วนที่แปลผิดพลาดแล้ว RI จะตรวจสอบข้อมูลที่เป็นผลลัพธ์ในส่วนของขั้นตอนการจัดการคำเบื้องต้น (Word Treatment) เพื่อให้แน่ชัดว่าการแปลภาษาทำได้ถูกต้อง จากนั้นทำการเรียงลำดับคำ (Word Ordering) ใหม่เพื่อจัดโครงสร้างแล้วจึงคืนค่าตัวแทนการแปลภาษาให้กับขั้นตอน TCE เพื่อวิเคราะห์ความหมายอีกที

ในกรณีที่ค่าในส่วน CONT หรือ QSTORE มีค่าไม่เท่ากัน ค่าของ SURFACE และ WORDASSO จะถูกส่งมาที่ RI ข้อมูลในส่วนของ SURFACE จะแสดงให้เห็นถึงคำที่ทำให้เกิดการ

แปลความหมายที่ผิด จากนั้น RI จะทำการเลือกคำให้ SURFACE ใหม่ โดยเลือกจากคำที่มีค่าตัวเลข WordAsso เหมือนกับใน WORDASSO

Example 5:

SL: The big ugly duckling likes Joan.
TC: ลูกเป็ด ขี้เหร่ โต ตัว นั้น เหมือน โจแอน
(lûugpèd-duckling) (khîilèe-ugly) (too-big) (tua-clas) (nán-the) (myˇan-like) (joo?ɛn-Joan)

```
0 the 1 big 2 ugly 3 duckling 4 likes 5 joan 6

phrase
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE like
        QUANTS e_list
. . .
```

```
0 ลูกเป็ด 1 ขี้เหร่ 2 โต 3 ตัว 4 นั้น 5 เหมือน 6 โจแอน

phrase
CONT psoa
        NUCLEUS sem_like
            WORDASSO 13-18-1
            LIKED [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKENER [0]
            SURFACE เหมือน
        QUANTS e_list
. . .
```

FIGURE 9. The CONT values of the SL and that of the TL parses.

จากภาพข้างต้น (ภาพที่ 9) จะเห็นได้ว่าข้อมูลในส่วน CONT ของภาษาตั้งต้นและตัวแทนการแปลภาษาต่างกันอย่างเห็นได้ชัด เรารู้ได้ว่าคำไหนผิดจาก SURFACE ซึ่งในที่นี้คือคำว่า like ซึ่งมีความหมายว่า “ชอบ” จัดอยู่ในหมวด 2-6-2-4 RI จึงเข้าไปดูในตารางผลลัพธ์ที่ได้จากขั้นตอนการจัดการคำเบื้องต้น (Word Treatment) ในตารางที่ 4 แล้วจึงเลือกคำว่า “ชอบ” ออกมา (ด้วยเหตุผลว่ามี WordAsso เหมือนกัน) หลังจากนั้นจึงทำขั้นตอนการเรียงลำดับคำ (Word Ordering) อีกครั้ง ภาพที่ 11 แสดงให้เห็นถึงข้อมูลของความหมายใน SL และ TC ที่ผ่านการแก้ไขแล้ว จะเห็นว่าข้อมูลที่ได้ไม่แตกต่างกันแล้วจึงสรุปว่าการแปลภาษาที่ได้เหมาะสมแล้ว

TABLE 4. A Part of Word Treatment Output.

| English | Word Treatment Output | Description |
|---|---|---|
| like | ชอบ (chəˆə), 2-6-2-4<br>เหมือน (myˇan), 13-18-1<br>คล้าย (khláaj), 13-18-2 | to regard with pleasure or fondness<br>in the same way as; with the same quality as<br>similar to; almost but not exactly the same |

TC: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) เหมือน (myˇan-like) โจแอน (joo?ɛn-Joan)

TC with the new selection: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) ชอบ (chəˆə-like) โจแอน (joo?ɛn-Joan)

Repaired TC: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) ชอบ (chəˆə-like) โจแอน (joo?ɛn-Joan)

FIGURE 10. The repaired TC of Example 5.

```
0 the 1 big 2 ugly 3 duckling 4 likes 5 joan 6

phrase
QSTORE ne_set_quant
    ELT quant
      DET the
      RESTIND nom_obj
          INDEX [0] ref
              IGENN gend
              INUM sg
              IPER per
          MODE reff
          RESTR ne_set_psoa
              ELT psoa
                NUCLEUS size
                    WORDASSO 13-8-1
                    INSTANCE [0]
                    SURFACE big
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [0]
                      SURFACE ugly
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [0]
                        SURFACE duckling
                    QUANTS e_list
                  ELTS e_set
    ELTS e_set
. . .
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE like
        QUANTS e_list
CONX conx
        BACKGR ne_set_psoa
           ELT psoa
             NUCLEUS naming
                 WORDASSO wordasso
                 BEARER [1]
                 NAME joan
                 SURFACE surface
             QUANTS e_list
           ELTS e_set
. . .
```

```
0 ลูกเป็ด 1 ขี้เหร่ 2 โต 3 ตัว 4 นั้น 5 ชอบ 6 โจแอน 7

phrase
QSTORE ne_set_quant
    ELT quant
      DET the
      RESTIND nom_obj
          INDEX [0] ref
              IGENN gend
              INUM sg
              IPER per
          MODE reff
          RESTR ne_set_psoa
              ELT psoa
                NUCLEUS size
                    WORDASSO 13-8-1-1
                    INSTANCE [0]
                    SURFACEโต
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [0]
                      SURFACE ขี้เหร่
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [0]
                        SURFACE ลูกเป็ด
                    QUANTS e_list
                  ELTS e_set
    ELTS e_set
. . .
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE ชอบ
        QUANTS e_list
CONX conx
        BACKGR ne_set_psoa
           ELT psoa
             NUCLEUS naming
                 WORDASSO wordasso
                 BEARER [1]
                 NAME โจแอน
                 SURFACE surface
             QUANTS e_list
           ELTS e_set
. . .
```

FIGURE 11. The semantic information of the SL and that of the repaired TC.

**5. บทสรุป**

GRMT ประกอบไปด้วย 3 ขั้นตอนก็คือ ALMT TCE และ RI ในขั้นตอน ALMT จะสร้างตัวแทนการแปลภาษาที่เรียบง่ายขึ้นมาตรงๆ หลังจากนั้นในขั้นตอน TCE จะทำการวิเคราะห์ตัวแทนการแปลภาษาที่ได้จากขั้นตอน ALMT ว่าถูกต้องตามหลักโครงสร้างไวยากรณ์หรือยัง แล้วจึงตรวจสอบว่าตัวแทนการแปลภาษายังคงความหมายเดิมจากภาษาตั้งต้นหรือเปล่า ถ้าไม่จะมีการซ่อมแซมและแก้ไขในขั้นตอน RI จากนั้นส่งตัวแทนการแปลภาษาที่แก้ไขแล้วกลับไปยังขั้นตอน TCE เพื่อให้ตรวจสอบอีกครั้ง ซึ่งทั้งสองขั้นตอนนี้จะช่วยเน้นย้ำความถูกต้องของการแปล

GRMT ประมวลผลแยกภาษาตั้งต้นกับภาษาที่ต้องการออกจากกันโดยสิ้นเชิง เพราะว่า GRMT นั้นเห็นว่าภาษาแต่ละภาษานั้นมีข้อแตกต่างอยู่มาก แต่เราสามารถจัดลักษณะพิเศษของแต่ละภาษาให้เป็นกลุ่มๆ ได้ ซึ่งทำให้การแปลภาษาง่ายขึ้นมาก

# เอกสารอ้างอิง

[1] Kanlaya Naruedomkul, Nick Cercone . Generate and Repair Machince Translation. Computational Intelligence, Volume 18, Number 3, 2002.

[2] Pollard, C., and I. A. Sag.1994. Head-Driven Phrase Structure Grammar.Center for the Language and Information, Stanford. The University of Chicago Press, Chicago, London.

[3] Sornlertlamvanich ,V., W.Pantachat, and S. meknavin. 1994. Classifier Assignment Corpus-Based Approach. The Fifteenth Computational Linguistic Symposium (COLING-94), Kyoto, Japan, vol. 1, pp.556-561

*Computational Intelligence*, Volume 18, Number 3, 2002

# GENERATE AND REPAIR MACHINE TRANSLATION

KANLAYA NARUEDOMKUL
*Department of Mathematics, Mahidol University, Bangkok, Thailand*

NICK CERCONE
*Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada*

We propose Generate and Repair Machine Translation (GRMT), a constraint-based approach to machine translation that focuses on accurate translation output. GRMT performs the translation by generating a *Translation Candidate* (TC), verifying the syntax and semantics of the TC and repairing the TC when required. GRMT comprises three modules: Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE) and Repair and Iterate (RI). The key features of GRMT are simplicity, modularity, extendibility, and multilinguality.

An English–Thai translation system has been implemented to illustrate the performance of GRMT. The system has been developed and run under SWI-Prolog 3.2.8. The English and Thai grammars have been developed based on Head-Driven Phrase Structure Grammar (HPSG) and implemented on the Attribute Logic Engine (ALE). GRMT was tested to generate the translations for a number of sentences/phrases. Examples are provided throughout the article to illustrate how GRMT performs the translation process.

*Key words*: Generate and Repair Machine Translation, machine translation, multilingual machine translation.

## 1. INTRODUCTION

In 1996, we proposed Generate and Repair Machine Translation (GRMT) (Naruedomkul and Cercone 1997). GRMT is designed to increase the accuracy and efficiency of machine translation (MT). The GRMT architecture is designed to take advantage of, and have advantage over, the direct, transfer, interlingual, and nonlinguistic approaches to MT with respect to several translation aspects: simplicity, accuracy, and multilingualism. GRMT integrates the best features of each approach. The GRMT process is relatively simple and straightforward, not unlike the direct method. However, GRMT is more concerned with preserving linguistic information to produce an accurate translation result, like the transfer approach. GRMT also treats the source language (SL) and target language (TL) separately for easy management in multilingual MT systems, like the interlingual approach.

Unlike the transfer or interlingual approaches, GRMT generates a translation candidate (TC) directly from the input to avoid information loss during the transfer process of the transfer approach and to avoid difficulties in defining neutral concepts for different languages in the interlingual approach (Cercone and Naruedomkul 1997). To ensure that the TC can be generated quickly, simply, and efficiently, GRMT generates the TC by considering the differences between language pairs in terms of syntax and semantics without performing any sophisticated analysis. GRMT analyzes the TC to verify its accuracy. The TC will be repaired, if required, according to the diagnosis which is indicated in the analysis stage. Subsequently, the repaired TC will be analyzed to determine if it still has a sufficiently different meaning from the SL. The analysis and repair processes iterate until the TC conveys the same meaning as the SL. These two stages ensure the accuracy of the translation result. Based on these notions, GRMT (Figure 1) is composed of three phases: Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE), and Repair and Iterate (RI).

GRMT takes into account the differences between languages in a unique way, hence a further advantage. If languages can be grouped according to various characteristics—for example, plurality, tenses, passive voice, etc.—which they have in common, then the translation between groups can be performed more simply by GRMT. For example, Group 1 consists of English, French, and Spanish; Group 2 consists of Chinese, Japanese, and Thai. To perform the translation between these two groups, the transfer approach requires six SL

© 2002 Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 108 Cowley Road, Oxford, OX4 IJF, UK.

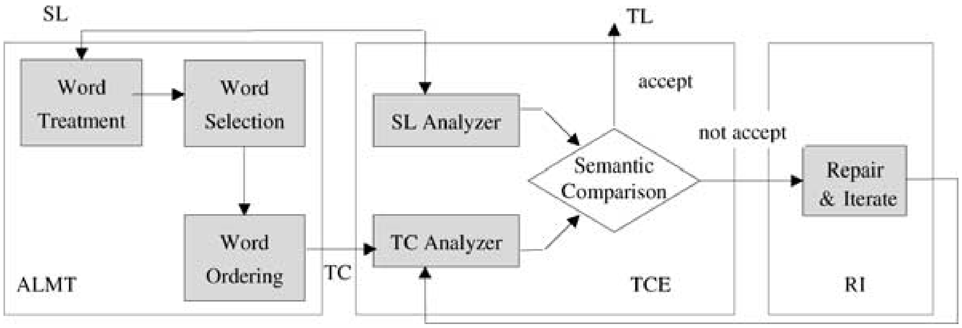


FIGURE 1. GRMT Architecture.

analyzers, six TL generations, and 18 sets of transfer rules, whereas GRMT requires six analyzers and two sets of constraint applications.

An English–Thai translation system has been developed based on the GRMT approach. Our initial experiments of ALMT (Naruedomkul, Cercone, and Sirinaovakul 1999) indicate that the translation candidates can be generated quickly with relatively high accuracy. We have greatly expanded the dictionaries used in the earlier experiment fivefold. All necessary knowledge bases required were updated. ALMT has been improved to increase the performance of GRMT. The English–Thai translation system we developed has been used to translate a number of sentences and phrases. These sample sentences and phrases were collected from different sources: newspapers, textbooks, articles, fairy tales, and fortune cookies. This article presents the design and some experiments of GRMT.

## 2. ANALYSIS LITE MACHINE TRANSLATION

To generate an appropriate translation candidate from the source language, ALMT performs generation in three phases: word treatment, word selection, and word ordering.

*Word Treatment*. Two steps are performed in this phase: source language constraints application and dictionary look-up (Naruedomkul and Cercone 1997). The SL constraints are applied to narrow the scope of possible TL words that correspond to each SL word. Dictionary look-up maps all corresponding words in the TL to each SL word.

SL constraints are characteristics of the SL which differ from those of the TL. Some constraints that are considered in the English–Thai translation system, e.g., plurality,[1] continuous tense, passive voice, and negation, are illustrated in Table 1.

Since Thai word form does not have an inflection resulting from plurality or tense, the inflections of plural noun and present participle forms in English are removed in this step. Once the inflections are removed, the features *plural* and *ing* are added to preserve the meaning of the original sentence. The structure of a passive voice is examined as well; the verb inflection and auxiliary *to be* are removed. The feature *passive* is added to indicate the passive voice. The features *plural*, *ing*, and *passive* will be replaced with appropriate corresponding TL words in the ordering step of ALMT.

[1]This constraint does not apply in the case of the irregular plural forms, e.g., children, feet, men, women, and so on. The irregular plural forms are entries in the dictionaries.

TABLE 1. Some SL Constraints considered in the English–Thai MT system

| SL Constraints | Descriptions | Examples |
|---|---|---|
| Plurality | noun_(e)s → noun + plural | The books . . . → The book + plural |
| Continuous tense | V to be + V_ing → V + ing | I am swimming. → I + ing + swim |
| Passive voice | be + V3 → passive + V | He was arrested. → He passive + arrest |
| Negative | do + not → not | He does not eat → He not eat |

The auxiliary *to do* which precedes a negative *not* is discarded because it is not required in the same expression in Thai. Figure 2 illustrates the outputs of SL constraints applications on Example 1. The inflection $-s$ of the word *symbol* in Example 1 is removed and the feature *plural* is added, the passive voice form *are used* is replaced with *passive use*, the auxiliary *do* is discarded, and the continuous tense form is replaced with *ing talk*.

After the SL constraints are applied, each SL word will be used as a keyword to search for its corresponding words in the TL. If the keyword used is found in the SL-TL dictionary, all possible corresponding TL words will be attached to that SL word. If the keyword is not found, inflectional analysis is performed before searching again. The inflections of each entry are not entries in the dictionaries; therefore the size of the SL dictionary and the search times are reduced. An output of this step is a list of all possible corresponding words of SL words in TL with their WordAsso numbers.

*Word Selection*. The criterion used in the word selection process is the semantic relationship between words. The semantic relationship we developed is a link between word association numbers (WordAsso) (Naruedomkul and Cercone 1999). WordAsso is a number assigned to each word class. WordAsso is our modification to the "A Kind Of" (AKO) information which was designed to be used in "The research and development cooperation project on a machine translation system for Japan and its neighboring countries" (CICC 1995), the topic hierarchy for physical objects (Schubert, Goebel, and Cercone 1979), and Hypernym in WordNet 1.6 (Cognitive Science Laboratory 1997). We then further developed our notion of word classification by classifying words into categories based on the characteristics which they share. We were concerned with consistency as well. The improved WordAsso results in more accurate word selections than the earlier version reported in Naruedomkul and Cercone (1997).[2] In addition, we classify words in any language under the "same umbrella." Words must be classified according to the same criteria, regardless of language; therefore, the GRMT classification can be applied to a multilingual MT system. The details of the word selection process can be found in Naruedomkul and Cercone (1997).

Example 1:
SL: Algebraic <u>symbols</u> <u>are used</u> when you <u>do not</u> know what you <u>are talking</u> about.
SL Cons- output: Algebraic <u>plural</u> symbol <u>passive use</u> when you <u>not</u> know what you <u>ing talk</u> about.

FIGURE 2. SL constraints applications examples.

[2]In that version, Word Association Number is refereed to as the AKO number.

*Word Ordering*. There are two steps in this phase, word addition and word ordering. The syntactic level differences between SL and TL are reconsidered at this point to complete the TC. Some words that are necessary in the target language, not only to retain the meaning of source language, but also to make them grammatically correct, are added into the string before the ordering can be performed. In the case of noninflection languages, e.g., Thai, Chinese, Japanese, etc., the *past tense* and the *plurality* are expressed by additional words. In the case of plurality, a *classifier* is also required. Classifiers are used to express not only a quantity but also to modify a noun (Sornlertlamvanich, Pantachat, and Meknavin 1994). Each noun relates to a specific classifier; therefore, the classifier relation was designed in the form of WordAsso to be used in selecting the appropriate classifier for each noun. Details and examples can be found in Naruedomkul and Cercone (1999).

After all necessary words have been added, ALMT rearranges the selected words in grammatical order according to the ordering rule without performing any analysis. The structure of the Thai phrase is similar to that of the English phrase in that the typical sentence contains subject, verb, and object, in that order. However, some structures are different, e.g., in Thai the head (noun) must precede its attributes, possessive pronouns, and determiners. The negative of *can*, *could*, and *must* is formed in the reverse order of the order in English.

Table 2 presents the results of applying ALMT to Example 2. The SL of Example 2 is shown in the first column. None of constraints are applied in Example 2 as illustrated in the second column. Each word in the second column is used as a keyword to search for the corresponding words in Thai. All possible meanings of each SL word are shown in the third column. Some of them have more than one meaning, e.g., *old*, *in*, *live*, *with*, and so on. The appropriate meaning of *old* and *in* can be selected by considering the semantic relationship between words and the choice for each of these is shown in the fourth column. However, appropriate words for *live* and *with* cannot be selected in the same manner because there is no explicit relationship between these words and words in their proximity (according to the WordAsso relationship template). Therefore, the first meaning appearing in the list of meanings for each word is selected. All selected words are shown in the fourth column. The word ได้ (dâj), the fifth column, is added to clarify the past tense (lived). The classifiers คน (khon), หลัง (laù$\eta$), and ตัว (tua) are also added according to Thai grammar. The indefinite determiners *a* and *an* in this expression correspond to the word หนึ่ง (ny`$\eta$) in Thai indicate the need for classifiers for the words ผู้หญิง (phûujiù$\eta$-woman), แมว (m$\varepsilon\varepsilon$w-cat) and ไก่ (kàj-hen), respectively. The word ผู้หญิง (phûujiù$\eta$-woman) belongs to the class Female (1-1-1-1-1-2), a subclass of Human (1-1-1-1). A noun that belongs to the class 1-1-1-1 is compatible with a classifier with the WordAsso

Example 2: An old woman lived in the cottage, with a fat black cat and a plump brown hen.

TC: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สีดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง[3]
CT: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สีดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง
(phûujiˇ$\eta$[4]-woman) (k$\varepsilon$\`$\varepsilon$-old) (khon-clas) (ny\`$\eta$-an) (dâj-past) (ju\`u-live) (naj-in) (krath$\partial$ˆm -cottage) (laˇ$\eta$-clas) (nán-the) (kàb-with) (m$\varepsilon\varepsilon$w-cat) (siˇidam-black) (?ûan-fat) (tua-clas) (ny\`$\eta$-a) (l$\varepsilon$\`-and) (kàj-hen) (siˇinámtaan-brown) (?ùab-plump) (tua-clas) (ny\`$\eta$-a)

[3]Thai language is written in a string of words with no explicit boundary marker;therefore, the different segmentations result in the different meanings or no meaning. In this article, awhite space is used to specify the word boundary to make it clear to the reader.

[4]Phonetic transcription of Thai provided in Thai–EnglishStudent's Dictionary compiled by Mary R. Haas (Hass 1964).

TABLE 2. ALMT Steps Applied to the Sentence of Example 2.

| English | Constraint Application Output | Word Treatment Output | Selected Word in Thai | Word Addition | Word Ordering |
|---|---|---|---|---|---|
| An | An | หนึ่ง (ny`η) | หนึ่ง (ny`η) | | ผู้หญิง (phûuji˘η) |
| | | | | คน (khon) | แก่ (kε`ε) |
| old | old | แก่ (kε`ε), เก่า (kàw) | แก่ (kε`ε), | | คน (khon) |
| woman | woman | ผู้หญิง (phûuji˘η) | ผู้หญิง (phûuji˘η) | | หนึ่ง (ny`η) |
| | | | | ได้ (dâj) | ได้ (dâj) |
| lived | live | อยู่ (ju`u), มีชีวิต (miichiiwíd), ดำรงชีวิต (damroηchiiwíd) | อยู่ (ju`u) | | อยู่ (ju`u) |
| in | in | ใน (naj), เข้ามา(khâwmaa), เข้าไป(khâwpaj), อยาง(jàaη) | ใน (naj) | | ใน (naj) |
| the | the | นั้น (nán) | นั้น (nán) | | กระท่อม (krath∂^m) |
| | | | | หลัง (la˘η) | หลัง (la˘η) |
| cottage | cottage | กระท่อม (krath∂^m) | กระท่อม (krath∂^m) | | นั้น (nán) |
| with | with | กับ (kàb), ด้วย (dûaj), ซึ่งมี (sy^η) | กับ (kàb) | | กับ (kàb) |
| a | a | หนึ่ง (ny`η) | หนึ่ง (ny`η) | | แมว (mεεw) |
| | | | | ตัว(tua) | สีดำ (si˘idam) |
| fat | fat | อ้วน (?ûan), ไขด้น (kha˘jman) | อ้วน (?ûan) | | อ้วน (?ûan) |
| black | black | ดำ (dam), สีดำ (si˘idam) | สีดำ (si˘idam) | | ตัว (tua) |
| cat | cat | แมว (mεεw), ผู้หญิวนม่มั (phûuji˘ηmâjdìi) | แมว (mεεw) | | หนึ่ง (ny`η) |
| and | and | และ (lε`) | และ (lε`) | | และ (lε`) |
| a | a | หนึ่ว (ny`η) | หนึ่ว (ny`η) | | ไก่ (kàj) |
| | | | | ตัว (tua) | สีน้ำตาล (si˘inámtaan) |
| plumb | plumb | อวบ (?ùab) | อวบ (?ùab) | | อวบ (?ùab) |
| brown | brown | สีน้ำ ตาล(si˘inámtaan), เกรียม(kriam), คล้ำ (khlám) | สีน้ำ ตาล(si˘inámtaan) | | ตัว (tua) |
| hen | hen | ไก่ (kàj) | ไก่ (kàj) | | หนึ่ง (ny`η) |

```
clf_rel('1-1-1-1',[[wasso('คน',['2-4-2-1-1-1'])]]).
clf_rel('1-1-1-2',[[wasso('ตัว',['2-4-2-1-1-2'])]]).
clf_rel('1-1-2-1-6-1',[[wasso('แห่ง',['2-4-2-1-2-2'])]]).
clf_rel('2-4-3-1-2',[[wasso('วัน',['2-4-2-3'])]]).
clf_rel('1-1-2-1-1-2-2',[[wasso('หลัง',['2-4-2-1-2-10'])]]).
```

FIGURE 3. Examples of classifier relations.

number 2-4-2-1-1-1 based on the classifier relation illustrated in Figure 3. Therefore, the classifier คน (khon) with 2-4-2-1-1-1 is selected for the word ผู้หญิง (phûujiù$\eta$-woman). The words แมว (m$\varepsilon\varepsilon$w-cat) and ไก่ (kàj-hen) belong, respectively, to the classes mammal (1-1-1-2-1-1) and fowl (1-1-1-2-1-2-2). Both are subclasses of animal (1-1-1-2). Because a noun with WordAsso number 1-1-1-2 relates to a classifier with 2-4-2-1-1-2 according to the classifier relation shown in Figure 3, the classifier ตัว (tua) with 2-4-2-1-1-2 is selected for the words แมว (m$\varepsilon\varepsilon$w-cat) and ไก่ (kàj-hen).

The definite determiner *the* corresponds to the word นั้น (nán) in Thai and indicates the need for classifiers for the word กระท่อม (krath$\partial$^m-cottage). The word กระท่อม (krath$\partial$^m-cottage) belongs to the class Housing (1-1-2-1-1-2-2) which is compatible with a classifier with the WordAsso number 2-4-2-1-2-10 based on the classifier relation illustrated in Figure 3. Therefore, the classifier หลัง (la$\check{}\eta$) with the WordAsso 2-4-2-1-2-10 is selected for the word กระท่อม (krath$\partial$^m-cottage).

The selected words in each noun phrase, *an old woman*, *the cottage*, *a fat black cat*, and *a plump brown hen* are rearranged into Thai grammatical order as illustrated in the last column of Figure 3. The generated TC for Example 2 is exactly the same as the *Correct Translation* (CT).

## 3. TRANSLATION CANDIDATE EVALUATION

The second phase of GRMT, TCE, analyzes the generated translation candidate to determine whether the TC retains the meaning of the source language. TCE analyzes both the SL and the TC in parallel, then compares the parses semantically alone, because there are syntactic level differences between languages. If the semantic results are the same, the TC will be deemed an appropriate translation. If the semantic results are different, the TC will be repaired in the third phase, Repair and Iterate. TCE comprises two modules: the analyzer and semantic comparison.

*The Analyzer*. The analysis module analyzes the TC to examine its syntax and semantics: whether the TC is grammatically correct according to the TL grammar and whether the TC retains the meaning of the SL. Therefore, two steps are performed by the analyzer, *Parsing* and *Semantic Extraction*.

*Parsing* is applied to both the SL and the TC. In our implementation, we have developed grammars for English and Thai based on Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag 1994). Our grammar is a modification of the grammars that were provided by Gerald Penn (Penn 1994) and Colin Matheson (Matheson 1996). The grammars we developed have been implemented using the Attribute Logic Engine (ALE) version 3.2 Beta. ALE is an integrated phrase structure parsing and definite clause logic programming system in which the

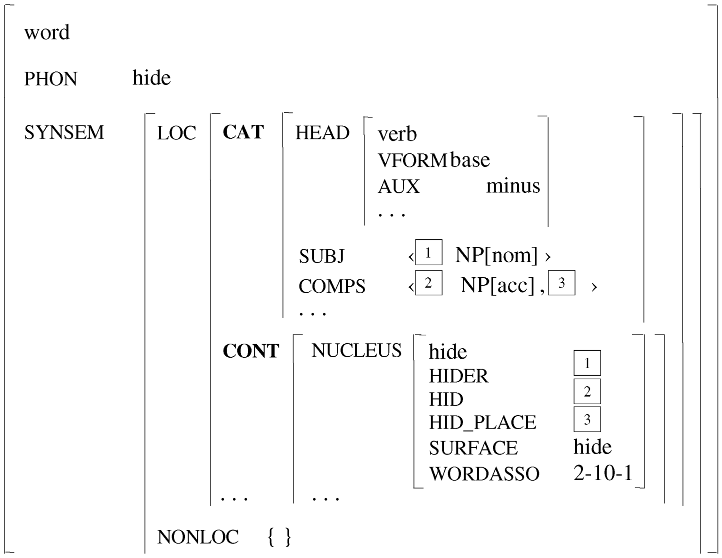


FIGURE 4. The lexical entry for *hide.*

terms are typed as feature Structures (Carpenter and Penn 1999). Key advantages of HPSG are the rich information lexicon, small number of grammar rules, and the structure sharing. HPSG views the grammatical categories as complexes with internal structure which allow us to say that the two categories are the same in certain aspects, while remaining different in others. Therefore, HPSG gives each word a category built up from a set of features called the Feature Structure. Feature Structure is a description of an object; it specifies some or all of the information that is asserted to be true of the object. The features CATEGORY (CAT) and CONTENT (CONT) represent the syntactic and semantic information of an object, respectively. Figure 4 represents a partial description of the lexical entry *hide* in an attribute-value matrix (AVM) diagram.[5] This representation indicates that *hide* is a word.[6] Its syntactic category is a verb in uninflected form (verb, VFORM:base). It is not an auxiliary verb (AUX:minus). A subject of *hide* is a (nominal) noun phrase[7] (SUBJ:$NP_{[nom]}$) and this subject must unify with the CONT value of *hide*, HIDER (indicated by index 1). The first complement of *hide* is an accusative noun (COMPS:$NP_{[acc]}$) and it must unify with the CONT value of *hide*, HID. The second complement must unify with the CONT value of *hide*, HID_PLACE. Figure 5 illustrates the parse of the SL in Example 3 and the parse of its TC. The TC is generated by ALMT. The parses are in ALE representation. Both SL and TC are licensed by our grammars. Their syntax are shown in the dotted box. Further details on grammar rules applications can be found in Naruedomkul (2000).

*Semantic Extraction* extracts the semantic information of the SL and the TC from their parses so that we can cross-examine the meaning of the TC with that of the SL. The semantic

[5] AVM is the standard method of representing grammatical information in modern computation grammar theories.

[6] Utterances in HPSG are modeled in terms of feature structures of type *sign*, with its two immediate subtypes *word* and *phrase* (Ginzburg and Sag 1998).

[7] For English, nouns are classified into nominative (nom) and accusative (acc).

Example 3: The ugly duckling hides his head under his wing.

SL: The ugly duckling hides his head under his wing.

TC: ลูกเป็ด ขี้เหร่ ตัว นั้น ซ่อน หัว ของเขา ใต้ ปีก ของเขา
(lûugpèd-duckling) (khîirèe-ugly) (tua-clas) (nán-the) (s∂ˆ∂n-hide) (huˇa-head)
(kh∂ˇ∂ηkhaw-his) (tâaj-under) (pìig-wing) (kh∂ˇ∂ηkhaw-his)

```
STRING:
0 the 1 ugly 2 duckling 3 hides 4 his 5 head
6 under 7 his 8 wing 9
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                         NUCLEUS beauty
                               WORDASSO 13-8-4
                               INSTANCE [4]
                               SURFACE ugly
. . . .
SYNSEM synsem
     LOC loc
        CAT cat
           COMPS e_list
           HEAD verb
              AGR agr
                 GENN gend
                 NUM num
                 PER per
              AUX minus
              INV minus
              LANG eng
              MOD none
              NEG boolean
              PRD boolean
              VFORM fin
           MARKING unmarked
           SPR list_synsem
           SUBJ e_list
        CONT psoa
           NUCLEUS hide2
                 WORDASSO 2-10-1
                 HID1 [1]
                 HID_PLACE [3]
                 HIDER1 [4]
                 SURFACE hide
           QUANTS e_list
        CONX conx
. . .
```

```
STRING:
0 ลูกเป็ด 1 ขี้เหร่ 2 ตัว 3 นั้น 4 ซ่อน 5 หัว
6 ของเขา 7 ใต้ 8 ปีก 9 ของเขา 10
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                         NUCLEUS beauty
                               WORDASSO 13-8-4
                               INSTANCE [4]
                               SURFACE ขี้เหร่
. . .
SYNSEM synsem
     LOC loc
        CAT cat
           COMPS e_list
           HEAD verb
              AGR agr
                 GENN gend
                 NUM num
                 PER per
              AUX minus
              INV minus
              LANG thai
              MOD none
              NEG boolean
              PRD boolean
              VFORM bse
           MARKING unmarked
           SPR ne_list_synsem
              HD synsem
                 LOC loc
                    CAT cat
                       COMPS list_synsem
                       . . .
                       MARKING comp
                       SPR list_synsem
                       SUBJ list_synsem
. . .
           SUBJ e_list
        CONT psoa
           NUCLEUS hide2
                 WORDASSO 2-10-1
                 HID1 [1]
                 HID_PLACE [3]
                 HIDER1 [4]
                 SURFACE ซ่อน
           QUANTS e_list
        CONX conx
. . .
```

FIGURE 5. The syntax of Example 3 and that of its translation.

representation of an expression in GRMT is generally based on the representation provided by Pollard and Sag (1994) and Sag and Wasow (1999). The features CONT, CONX and QSTORE hold the semantics of the object. The semantic representation described in CONT value of *hide* (Figure 4) corresponds to the situation in which[1] (*the hider*) hides[2] (*the hid)* in (or under, from, etc.)[3] (*the hid_place*). The features SURFACE and WORDASSO indicate the word form and WordAsso number of the lexical entry. The features SURFACE and WORDASSO are used in the Repair and Iterate phase. The feature QSTORE is a storage for the quantifiers. The feature CONX contains linguistic information that bears on certain context-dependent aspects of semantic interpretation. Examples of QSTORE and CONX values can be found in Figure 11. Further details can be found in Pollard and Sag (1994).

*Semantic Comparison*. In comparing the semantics between the SL and its TC, the values of the features CONT, QSTORE, and CONX are considered. If the values of these features of the parsed SL are the same as those of the parsed TC, TCE concludes that the TC does not require repair. If any of these features are different, TCE will provide the information of the parsed SL which differs from that of the TC parse. This information will be used in the next phase, Repair and Iterate. The comparison process begins by investigating the CONT value of both SL and TC parses. Figures 6 and 7 illustrate the simplification representation of semantic information of the SL and that of the TC in Example 3. The CONT values of both parses represent the same kind of relation involved, *hide2*, and the same types of persons or things who/which are participating in this relation, HIDER, HID, and HID_PLACE. Once the CONT values of the SL parse and that of the TL parse are recognized as the same, TCE then cross-examines the QSTORE and the CONX values of the SL parse with those of the TC parse to determine whether each variable (indices) in the CONT value is associated to the same object. The index[1] in the CONT value of both SL and TC parses is associated with a *bodypart* in the class of 1-4-1-1-1 (in the QSTORE value), it is *singular* and referred to as *head* in the SL and as หัว (huˇa) in the TL. หัว (huˇa) is the translation of the word *head* in Thai. The index[2] corresponds to a *bodypart* in the class of 1-4-1-1-3-1, it is singular and referred to as *wing* in the SL and ปีก (pìig) in the TL. Again, ปีก (pìig) is a translation of *wing* in Thai. The last index[3] is associated with an animal in the class of 1-1-1-2-1-2-1 which is *singular* and referred to as *duckling* in the SL and ลูกเป็ค (lûugpèd) in the TL. ลูกเป็ค (lûugpèd) is the translation of the word *duckling* in Thai. The CONX feature in this example contains no value. According to the comparison process, TCE found no difference between the CONT, QSTORE, and CONX of the SL parse and those of the TC parse. Therefore, the TC is deemed as an appropriate translation for the SL in the Example 3.

The CONT values of the SL parse and the TC parse in Example 4 (Figure 8) represent the same relation involved, *cover2*, and the same types of persons or things who/which are participating in this relation, COVERED, COVERED_PLACE and COVERER1. The SURFACE value in the CONT of the SL parse is *cover* and that of the TC is คลุม (khlum). The value of the feature SURFACE is a word form. คคลุม (khlum) is a translation of *cover*. However, their WORDASSO values are different, *cover* belongs to the class of 2-1-19 whereas คคลุม (khlum) belongs to the class of 2-1-19-2 as illustrated in Table 3. Even though *cover* and its translation คคลุม (khlum) carry the same meaning, they are different in the language usage. Differences in languages and cultures result in different circumstances of language usage. The word *cover* in *John covers his ears with his hands* corresponds to ปีค (pìd) in Thai. Therefore, when verifying the meaning of *cover*, the word and its translation are considered no different if they belong to the same superclass.

```
STRING:
0 the 1 ugly 2 duckling 3 hides 4 his 5 head 6 under 7 his 8 wing 9
CATEGORY:

phrase
QSTORE ne_set_quant
   ELT quant
     DET the
     RESTIND nom_obj
         INDEX [0] ref
            IGENN gend
            INUM num
            IPER per
         MODE mode
         RESTR ne_set_psoa
            ELT psoa
              NUCLEUS possess
                  WOROASSO wordasso
                  POSSESSED [1] ref
                       IGENN gend
                       INUM sg
                       IPER per
                  POSSESSOR [0]
                  SURFACE surface
              QUANTS e_list
            ELTS ne_set_psoa
              ELT psoa
                NUCLEUS bodypart
                    WORDASSO 1-4-1-1-1
                    INSTANCE [1]
                    SURFACE head
                        AGR agr
                          GENN gend
                          NUM num
                          PER per
                        LANG lang
                QUANTS e_list
              ELTS e_set
   ELTS ne_set_quant
     ELT quant
       DET the
       RESTIND nom_obj
           INDEX [2] ref
              IGENN gend
              INUM num
              IPER per
           MODE mode
           RESTR ne_set_psoa
              ELT psoa
                NUCLEUS possess
                    WOROASSO wordasso
                    POSSESSED [3] ref
                         IGENN gend
                         INUM sg
                         IPER per
                                        . . .
                    POSSESSOR [2]
                    SURFACE surface
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS bodypart
                      WORDASSO 1-4-1-1-3-1
                      INSTANCE [3]
                      SURFACE wing
                  QUANTS e_list
                ELTS e_set
     ELTS ne_set_quant
       ELT quant
         DET the
         RESTIND nom_obj
             INDEX [4] ref
                IGENN gend
                INUM sg
                IPER per
             MODE reff
             RESTR ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [4]
                      SURFACE ugly
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [4]
                        SURFACE duckling
                    QUANTS e_list
                  ELTS e_set
       ELTS e_set
. . .
CONT psoa
  NUCLEUS hide2
      WORDASSO 2-10-1
      HID1 [1]
      HID_PLACE [3]
      HIDER1 [4]
      SURFACE hide
  QUANTS e_list
CONX conx
  BACKGR e_set
```

FIGURE 6. The semantic information of the SL parse in Example 3.

## 4. REPAIR AND ITERATION

The idea of the RI phase is to perform the repair process if it is required and then return the repaired translation candidate to the TCE phase. TCE reanalyzes the repaired TC to determine if a different meaning from the source language remains.

```
STRING:
0 ลูกเปิด 1 ขี้เหร่ 2 ตัว 3 นั้น 4 ซ่อน 5 หัว 6 ของเขา 7 ใต้ 8 ปีก 9 ของเขา 10
CATEGORY:

phrase
QSTORE ne_set_quant
   ELT quant
     DET the
     RESTIND nom_obj
         INDEX [0] ref
            IGENN gend
            INUM num
            IPER per
         MODE mode
         RESTR ne_set_psoa
            ELT psoa
              NUCLEUS possess
                  WORDASSO wordasso
                  POSSESSED [1] ref
                       IGENN gend
                       INUM sg
                       IPER per
                  POSSESSOR [0]
                  SURFACE surface
              QUANTS e_list
            ELTS ne_set_psoa
               ELT psoa
                 NUCLEUS bodypart
                     WORDASSO 1-4-1-1-1
                     INSTANCE [1]
                     SURFACE หัว
                 QUANTS e_list
               ELTS e_set
   ELTS ne_set_quant
     ELT quant
       DET the
       RESTIND nom_obj
           INDEX [2] ref
              IGENN gend
              INUM num
              IPER per
           MODE mode
           RESTR ne_set_psoa
              ELT psoa
                NUCLEUS possess
                    WORDASSO wordasso
                    POSSESSED [3] ref
                         IGENN gend
                         INUM sg
                         IPER per
                    POSSESSOR [2]
                    SURFACE surface
                QUANTS e_list
              ELTS ne_set_psoa
                 ELT psoa
                   NUCLEUS bodypart
                       WORDASSO 1-4-1-1-3-1
                       INSTANCE [3]
                       SURFACE ปีก
                   QUANTS e_list
                 ELTS e_set
       ELTS ne_set_quant
         ELT quant
           DET the
           RESTIND nom_obj
               INDEX [4] ref
                  IGENN gend
                  INUM sg
                  IPER per
               MODE reff
               RESTR ne_set_psoa
                  ELT psoa
                    NUCLEUS beauty
                        WORDASSO 13-8-4
                        INSTANCE [4]
                        SURFACE ขี้เหร่
                    QUANTS e_list
                  ELTS ne_set_psoa
                    ELT psoa
                      NUCLEUS animal
                          WORDASSO 1-1-1-2-1-2-1
                          INSTANCE [4]
                          SURFACE ลูกเปิด
                      QUANTS e_list
                    ELTS e_set
         ELTS e_set
. . .
CONT psoa
   NUCLEUS hide2
      WORDASSO 2-10-1
      HID1 [1]
      HID_PLACE [3]
      HIDER1 [4]
      SURFACE ซ่อน
   QUANTS e_list
CONX conx
   BACKGR e_set
. . .
```

FIGURE 7. The semantic information of the TC parse in Example 3.

RI examines the result of TCE. The TCE output is the TC with the semantic information of the SL which differs from that of the TC. With this information, RI is able to detect the part of the TC which causes the mistranslation. The mistranslated part will be replaced with a more appropriate translation. RI searches the *Word Treatment output* for a more appropriate translation based on the information provided by TCE. The TC with the new selection is then

Example 4:

SL: John covers his head with his coat.
TC: จอห์น คลุม หัว ของเขา ด้วย เสื้อคลุม ของเขา
(cɔɔn-John) (khlum-cover) (huˇa-head) (khəˇəŋkhaw-his) (dûaj-with)
(syˆakhlum-coat)( khəˇəŋkhaw-his)

```
STRING:
0 john 1 covers 2 his 3 head 4 with 5 his 6 coat 7
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
        CONT psoa
            NUCLEUS cover2
                  WORDASSO 2-1-19
                  COVERED [1]
                  COVERED_PLACE [3]
                  COVERER1 [4] ref
                         IGENN masc
                         INUM sg
                         IPER per
                  SURFACE cover
            QUANTS e_list
        CONX conx
. . .
```

```
STRING:
0 จอห์น 1 คลุม 2 หัว 3 ของเขา 4 ด้วย 5 เสื้อคลุม 6 ของเขา 7
CATEGORY:

phrase
QSTORE ne_set_quant

. . .
        CONT psoa
            NUCLEUS cover2
                  WORDASSO 2-1-19-2
                  COVERED [1]
                  COVERED_PLACE [3]
                  COVERER1 [4] ref
                         IGENN masc
                         INUM sg
                         IPER per
                  SURFACE คลุม
            QUANTS e_list
        CONX conx
. . .
```

FIGURE 8. The CONT value of the SL parse and that of the TC of Example 4.

put through the *Word Ordering* module to revise its syntax. Once the revision is completed, the repaired TC is returned to TCE.

In the case that the CONT or the QSTORE value of the SL is different from that of the TC, the SL CONT value: SURFACE and WORDASSO features are passed to RI. The SURFACE value indicates the surface form of the word which causes the mistranslation. The WORDASSO value specifies the proper meaning of the word in question in terms of WordAsso number. Therefore, to repair the CONT or QSTORE value of the TC, RI reselects the corresponding word in the TL for the word specified in the SURFACE value. The reselection is done by searching the *Word Treatment Output* for the corresponding word which has the same WordAsso number as specified in the WORDASSO value.

In the case that the CONX value of the SL differs from that of the TC, the SL CONX value: the BEARER and NAME features are passed to RI. The BEARER value specifies the

TABLE 3. Class of *cover*.

| WordAsso and Class description | Word |
|---|---|
| 2-1-19 Cover as to place something on or over to protect or hide | cover |
| 2-1-19-1 Cover as to cover eyes, ears | ปิด (pìd) |
| 2-1-19-2 Envelop | ดคลุม (khlum) |

Example 5:

SL: The big ugly duckling likes Joan.
TC: ลูกเป็ด ขี้เหร่ โต ตัว นั้น เหมือน โจแอน
(lûugpèd- duckling) (khîilèe-ugly) (too-big) (tua-clas) (nán-the) (myˇan-like) (jooʔɛn-Joan)

```
0 the 1 big 2 ugly 3 duckling 4 likes 5 joan 6

phrase
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE like
        QUANTS e_list
. . .
```

```
0 ลูกเป็ด 1 ขี้เหร่ 2 โต 3 ตัว 4 นั้น 5 เหมือน 6 โจแอน

phrase
CONT psoa
        NUCLEUS sem_like
            WORDASSO 13-18-1
            LIKED [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKENER [0]
            SURFACE เหมือน
        QUANTS e_list
. . .
```

FIGURE 9. The CONT values of the SL and that of the TL parses.

index which associates with the certain name specified in the NAME value. RI repairs the TC by associating the right names to the right indices based on the information provided by TCE.

The CONT values of the SL and the TC parses of Example 5 are different as illustrated in Figure 9. The CONT value of the SL indicates that the word *like* is mistranslated (SURFACE: like) and its proper meaning in this expression is *to regard with pleasure or fondness* which is classified into the class of 2-6-2-4 (WORDASSO: 2-6-2-4). Therefore, RI begins the repair process by re-selecting the translation of the word *like.* RI searches the *WordTtreatment output* (Table 4) for the translation of *like* which has WordAsso number 2-6-2-4 and thus, the translation ชอบ (ch∂ˆ∂), is selected.

The word เหมือน (myˇan), in the TC is then replaced with ชอบ (ch∂ˆ∂). The TC with the new selection (Figure 10) is put through the *Word Ordering* module to revise its syntax. Once the repair processes successfully, the repaired TC is analyzed by TCE. Figure 11 illustrates the semantic information of the SL parse and that of the repaired TC. Their CONT, QSTORE, and CONX values are the same. The repaired TC is then deemed as an appropriate translation

TABLE 4. A Part of Word Treatment Output.

| English | Word Treatment Output | Description |
|---|---|---|
| like | ชอบ (ch∂ˆ∂), 2-6-2-4 | to regard with pleasure or fondness |
| | เหมือน (myˇan), 13-18-1 | in the same way as; with the same quality as |
| | คล้าย (khláaj), 13-18-2 | similar to; almost but not exactly the same |

TC: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) เหมือน (myˇan-like) โจแอน (joo?ɛn-Joan)

TC with the new selection: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) ชอบ (chəˆə-like) โจแอน (joo?ɛn-Joan)

Repaired TC: ลูกเป็ด (lûugpèd-duckling) ขี้เหร่ (khîilèe-ugly) โต (too-big) ตัว (tua-classifier) นั้น (nán-the) ชอบ (chəˆə-like) โจแอน (joo?ɛn-Joan)

FIGURE 10. The repaired TC of Example 5.

for the SL in Example 5. Further details of the repair process can be found in Naruedomkul (2000).

## 5. CONCLUDING REMARKS

GRMT consists of three phases: ALMT, TCE, and RI. ALMT generates the translation candidate in a simple, straightforward manner, similar to the direct approach, but ALMT is more efficient because it accounts for differences between language pairs in terms of both syntax and semantics and this analysis ensures that the generated TC is exact or close to the correct translation.

The TCE analyzes the TC to determine whether its syntax and semantics are grammatically correct and whether it conveys the meaning of the original sentence. If the TC does not, RI will repair it. These two stages, TCE and RI, ensure accuracy of the translation.

GRMT treats the SL and TL separately, as is also the case with the interlingual approach. GRMT is also aware of the differences between languages. Therefore, if languages can be grouped according to various characteristics, for example, plurality, continuous tenses, passive voice, etc., which they have in common, then the translation between groups can be performed more simply by GRMT.

Another aspect of concern in designing an MT system is the structure of the knowledge base, e.g., the constraints and the WordAsso information in the dictionary. The structure of each knowledge base component should be direct, intuitive, and easy to extend to a large-scale MT system. In generating a reliable TC, ALMT requires simple information. This simplicity ensures that knowledge bases required in the ALMT phase are easy to manage in a large-scale MT effort. The dictionary used in earlier reported experiments of ALMT (Naruedomkul and Cercone 1997) was expanded to 348 English words and 731 Thai words. The "semantic relationship" and the "ordering rule" were also updated. There has been no appreciable increase in processing time when running ALMT using this larger dictionary. The lexicons used in the TCE for both SL and TL are informative and easy to manage.

## ACKNOWLEDGMENTS

The authors are members of the Institute for Robotics and Intelligent Systems (IRIS) and wish to acknowledge the support of the Networks of Centres of Excellence Program of the Government of Canada, the Natural Sciences and Engineering Research Council (NSERC), and the participation of PRECARN Associates, Inc.

```
0 the 1 big 2 ugly 3 duckling 4 likes 5 joan 6

phrase
QSTORE ne_set_quant
    ELT quant
      DET the
      RESTIND nom_obj
          INDEX [0] ref
              IGENN gend
              INUM sg
              IPER per
          MODE reff
          RESTR ne_set_psoa
              ELT psoa
                NUCLEUS size
                    WORDASSO 13-8-1
                    INSTANCE [0]
                    SURFACE big
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [0]
                      SURFACE ugly
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [0]
                        SURFACE duckling
                    QUANTS e_list
                  ELTS e_set
    ELTS e_set
. . .
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE like
        QUANTS e_list
CONX conx
        BACKGR ne_set_psoa
            ELT psoa
              NUCLEUS naming
                  WORDASSO wordasso
                  BEARER [1]
                  NAME joan
                  SURFACE surface
              QUANTS e_list
            ELTS e_set
. . .
```

```
0 ลูกเป็ด 1 ขี้เหร่ 2 โต 3 ตัว 4 นั้น 5 ชอบ 6 โจแอน 7

phrase
QSTORE ne_set_quant
    ELT quant
      DET the
      RESTIND nom_obj
          INDEX [0] ref
              IGENN gend
              INUM sg
              IPER per
          MODE reff
          RESTR ne_set_psoa
              ELT psoa
                NUCLEUS size
                    WORDASSO 13-8-1-1
                    INSTANCE [0]
                    SURFACEโต
                QUANTS e_list
              ELTS ne_set_psoa
                ELT psoa
                  NUCLEUS beauty
                      WORDASSO 13-8-4
                      INSTANCE [0]
                      SURFACE ขี้เหร่
                  QUANTS e_list
                ELTS ne_set_psoa
                  ELT psoa
                    NUCLEUS animal
                        WORDASSO 1-1-1-2-1-2-1
                        INSTANCE [0]
                        SURFACE ลูกเป็ด
                    QUANTS e_list
                  ELTS e_set
    ELTS e_set
. . .
CONT psoa
        NUCLEUS like1
            WORDASSO 2-6-2-4
            LIKEE [1] ref
                IGENN fem
                INUM sg
                IPER per
            LIKER [0]
            SURFACE ชอบ
        QUANTS e_list
CONX conx
        BACKGR ne_set_psoa
            ELT psoa
              NUCLEUS naming
                  WORDASSO wordasso
                  BEARER [1]
                  NAME โจแอน
                  SURFACE surface
              QUANTS e_list
            ELTS e_set
. . .
```

FIGURE 11. The semantic information of the SL and that of the repaired TC.

## REFERENCES

CARPENTER, B., and G. PENN. 1999. ALE: The Attribute Logic Engine User's Guide Version 3.2 Beta. http://www.sfs.nphil.uni-tuebingen.de/~gpenn/ale.html#Obtain, accessed in May 2002.

CERCONE, N., and K. NARUEDOMKUL. 1997. Why GRMT? Vivek, **10**(3):12–15.

CICC. 1995. Thai Analysis Rules. Technical Report 6-CICC-MT46, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

COGNITIVE SCIENCE LABORATORY, PRINCETON UNIVERSITY. 1997. WordNet-A Lexical Database for English. http://www.cogsci.princeton.edu/~wn/, accessed in December 2002.

GINZBURG, J., and I. A. SAG. 1998. English Interrogative Constructures (Draft of July) *In* Construction: An HPSG Perspektive. *Edited by* I. A. SAG and A. KATHOL. Language Advanced Course, 10th European Summer School in Logic, Language and Information, Saarbrücken, 17–28 August.

HAAS, M. R. 1964. Thai–English Student's Dictionary. Stanford University Press, Stanford, CA.

MATHESON, C. 1996. HPSG in ALE. http://www.ltg.hcrc.ed.ac.uk/projects/ledtools/ale-hpsg/, accessed in April 2002.

NARUEDOMKUL, K. 2000. Beyond Intralingual: Squeezing More Accuracy into Machine Translation. PhD thesis (Draft), Department of Computer Science, University of Regina, Canada.

NARUEDOMKUL, K., and N. CERCONE. 1997. Steps Toward Accurate Machine Translation. *In* Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, NM, pp. 63–75.

NARUEDOMKUL, K., and N. CERCONE. 1999. The Role for Word Association Numbers in Machine Translation. *In* Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'99), Waterloo, Ontario, Canada, pp. 379–392.

NARUEDOMKUL, K., N. CERCONE, and B. SIRINAOVAKUL. 1999. English–Thai Translation: Initial Experiments with a Multiphase Translation System. Computational Intelligence, **15**(2):128–151.

PENN, G. 1994. hpsg.pl. http://www.sfs.nphil.uni-tuebingen.de/~gpenn/ale.html#AleGrammars, accessed in December 2002.

POLLARD, C., and I. A. SAG. 1994. Head-Driven Phrase Structure Grammar. Center for the Study of Language and Information, Stanford. The University of Chicago Press, Chicago, London.

SAG, I. A., and T. WASOW. 1999. Syntactic Theory: A Formal Introduction, CSLI Lecture Notes Number 92. Center for the Study of Language and Information, Stanford, CA.

SCHUBERT, L. K., R. G. GOEBEL, and N. CERCONE. 1979. The Structure and Organization of a Semantic Net for Comprehension and inference. *In* Associative Networks: Representation and Use of Knowledge by Computers. *Edited by* N. V. Findler. Academic Press, NY, pp. 121–175.

SORNLERTLAMVANICH, V., W. PANTACHAT, and S. MEKNAVIN. 1994. Classifier Assignment by Corpus-Based Approach. The Fifteenth Computational Linguistics Symposium (COLING-94), Kyoto, Japan, vol. 1, pp. 556–561.