ทรัพยากรเปิดทางภาษาสู่ความร่วมมือของการวิจัยและพัฒนา

The Open linguistic Resources CHanelled toward InterDisciplinary research (ORCHID)

ดร.วิรัช ศรเลิศล้ำวาณิช
virach@nwg.nectec.or.th

December 25, 1998

1 บทนำ

ORCHID เป็นแผนงานเพื่อสนับสนุนการร่วมกันสร้าง, ร่วมกันใช้, และร่วมกันพัฒนาทรัพยากรทางภาษาของภาษาไทย โดยมีจุดประสงค์หลักอยู่สองประการคือ เพื่อการแก้ปัญหาเรื่องกำแพงทางภาษา และการรักษาไว้เพื่อความคงอยู่ของภาษาและวัฒนธรรมไทย.

เราตระหนักดีถึงความสำคัญของภาษา ซึ่งนอกจากจะเป็นสื่อระหว่างคนกับคน แล้วยังเป็นรูปแทนความคิด และเป็นเครื่องมือในการใช้ความคิดด้วย. เครือข่ายคอมพิวเตอร์ในปัจจุบันทำให้ข้อมูลข่าวสารแพร่หลายไปอย่างรวดเร็ว. เครื่องมือที่ใช้ในการแสดงผล และการเตรียมข้อมูลข่าวสารนั้นจึงเป็นสิ่งจำเป็น. ด้วยเทคโนโลยีที่ก้าวหน้าไปอย่างรวดเร็ว, การที่เพียงจะสามารถแสดงได้ หรือป้อนข้อมูลได้นั้น ไม่เป็นที่เพียงพออีกแล้ว. การแสดงผลที่สวยงามถูกต้องตามแบบแผน หรือการเตรียมข้อมูลได้อย่างถูกต้องและรวดเร็วจึงเป็นสิ่งที่จำเป็นที่จะต้องพัฒนาให้ทันตามการเปลี่ยนแปลงของเทคโนโลยี.

ทรัพยากรทางภาษานอกจากจะเป็นแหล่งข้อมูลที่สำคัญแล้ว ยังเป็นปัจจัยที่สำคัญอันหนึ่งสำหรับ การศึกษาธรรมชาติของภาษา ซึ่งรวมถึงไวยากรณ์ คำศัพท์ และลักษณะของภาษา. ฉะนั้น *การรวบรวมข้อมูลอย่างมีระบบ* จึงมีความสำคัญยิ่ง. การศึกษาธรรมชาติของภาษาได้ก้าวหน้าไปมากพร้อมๆ กับพัฒนาการของระบบคอมพิวเตอร์และอัลกอริทึมในการคำนวณ. ความเอื้ออำนวยของเทคโนโลยีทางการประมวลผลทำให้เราสามารถศึกษาลักษณะของภาษาได้จากข้อมูลปริมาณมากๆ ได้ในเวลาอันรวดเร็ว. ผลที่ได้คือเราสามารถสรุปความรู้ทางภาษาจากข้อมูลจริงได้อย่างแหม่นยำ และครอบคลุม. ซึ่งผิดจากเมื่อในอดีตที่จำเป็นต้องพิจารณาจากความรู้ของตนเองเป็นหลัก, ศึกษาได้แต่ในวงแคบๆ และไม่สามารถตรวจสอบความถูกต้องได้อย่างครอบคลุม. ฉะนั้น *การรวบรวมข้อมูลที่มีปริมาณมากพอ และทันสมัย* จึงเป็นปัจจัยสำคัญอีกอย่างหนึ่ง.

จากแนวโน้มของพัฒนาการของเทคโนโลยีสารสนเทศ และการเตรียมพร้อมเพื่อการวิจัยและพัฒนาดังที่กล่าวมาข้างต้น, ทำให้ข้าพเจ้าได้คิดถึงแนวทางอันหนึ่งในการที่จะส่งเสริมโดยอาศัยมันสมองจากผู้รู้ผ่าน

เครือข่ายที่กำลังพัฒนาอยู่ในทุกวันนี้. ในบทความนี้ข้าพเจ้าจึงได้เสนอแนวคิดของ **ORCHID** ที่รวมถึงการเตรียมพร้อมของข้อมูล และการใช้งานเพื่อรองรับพัฒนาการของเทคโนโลยีสารสนเทศสู่อนาคต.

2 ข่าวสารบนเครือข่าย

ในปัจจุบัน, ข้อมูลข่าวสารสามารถแพร่หลายไปได้อย่างรวดเร็ว. ทั้งนี้ก็เนื่องมาจากสาเหตุที่ สำคัญสองประการคือ 1) ความแพร่หลายของระบบเครือข่าย ที่ทำให้บุคคลทั่วไปสามารถ เข้าถึงระบบได้โดยง่าย และ 2) พัฒนาการของเทคโนโลยีต่างๆ ที่ใช้บนระบบเครือข่าย ดังเช่น Web browser, ระบบสืบค้นข้อมูลบนระบบเครือข่าย, และอื่นๆ อีกมากมาย ที่ทำให้ สามารถเข้าถึงข้อมูลและข่าวสารได้อย่างแหม่นยำและรวดเร็ว. นับตั้งแต่ที่มีการกำกับข้อความ ด้วย HTML (HyperText Markup Language) [1] เพื่อบอกโครงสร้างทางตรรกศาสตร์ (logical structure) ของข้อความ. HTML เป็นภาษาที่แตกแขนงออกมาจากต้นตำรับของภาษาเพื่อการกำกับ (markup language) ที่รู้จักกันดีในชื่อของ SGML (Standard Generalized Markup Language) [2]. การใช้ภาษาเพื่อการ กำกับนี้จะทำให้ข้อความอิเล็กทรอนิกส์ (eletronic text) มีลักษณะพิเศษ คือ เป็นข้อความ ที่ไม่ขึ้นกับระบบจัดการ ซึ่งหมายความว่าทุกระบบ, ที่เข้าใจในมาตรฐานของภาษาเพื่อการ กำกับ, จะสามารถแสดงผลข้อมูลได้อย่างเหมาะสม. บางระบบอาจจะย่อหน้า หรือบางระบบ อาจจะขึ้นบรรทัดใหม่ เมื่อมีการกำกับย่อหน้า. แต่ละระบบอาจจะใช้ตัวอักษรขนาดต่างๆ กันใน การแสดงผลหัวข้อในระดับต่างๆ กัน. รูปที่ 1 แสดงตัวอย่างของการ แสดงผลที่แตกต่างกัน.

ภาษาเพื่อการกำกับนี้ จะเป็นตัวกำกับโครงสร้างของข้อความเพื่อให้ระบบสามารถจัดพิมพ์ หรือ แสดงผลได้ตามรูปแบบของต้นเอง. ปัจจุบันนี้ยังมีการกำหนดมาตรฐานของภาษาเพื่อการกำกับ ขึ้นมาใหม่, เรียกว่า XML (eXtensible Markup Language) [3]. XML เป็น subset ของ SGML, เป็นภาษาที่ออกแบบไว้เพื่อใช้ในการสื่อสารบนเครือข่าย โดยเฉพาะ. XML ต่างจาก SGML ตรงที่มีการคำนึงปัญหาต่างๆ ซึ่งอาจจะเกิดขึ้นได้ ในระหว่างการติดต่อผ่านเครือข่าย. XML จึงมีความยืดยุ่นมากในการกำกับ และผู้ใช้สามารถ กำกับให้รวบรวมข้อความหรือข้อมูลจากที่ต่างๆ พร้อมทั้งบอกลักษณะของข้อความหรือข้อมูลเหล่านั้นได้ด้วย. เหล่านี้ทำให้ XML เหมาะสำหรับการบรรทึกข้อมูล เนื่องจากว่ามีความยืดยุ่น พอที่จะสามารถอธิบายโครงสร้างทางตรรกศาสตร์ของข้อความต่างๆ ได้, ไม่ว่าจะเป็น แบบ ฟอร์ม (form), บรรทึก (memo), จดหมาย (letter), รายงาน (report), หนังสือ (book), สารานุกรม (encyclopedia), พจนานุกรม (dictionary) หรือฐานข้อมูล (database).

ผู้เขียนจึงขอสรุปไว้ในตอนท้ายของบทนี้ว่า ในการบรรทึกข้อมูลต่อไปในอนาคตนั้น, เรา จำเป็นต้องคำนึงถึงการใช้งานบนเครือข่าย. เครือข่ายที่พูดถึงตรงนี้ก็เป็นเครือข่ายสากล (Global Network, or World Wide Web) ไม่ใช่เครือข่ายท้องถิ่น (Local Area Network) อีกต่อไปแล้ว. การจัดเก็บข้อมูลที่เหมาะสมนั้นก็ควรจะต้องมีเนื้อหาของข้อความ (plain text) และข้อมูลของโครงสร้างทางตรรกศาสตร์กำกับไว้เพื่อที่จะให้ข้อมูลนั้นๆ เป็นอิสระจากอุปกรณ์ (device) และระบบ (system). การจัดเก็บข้อมูลในลักษณะนี้จะแตกต่างจากวิธีการเก็บแบบ WYSIWYG (what-you-see-is-what-you-get) ดังเช่นไฟล์ข้อมูลใน MS-Word เป็นต้น. การเก็บข้อมูลด้วยวิธีหลังนี้ค่อนข้างตรง, สะดวกต่อการแก้ไขและแสดงผล. แต่เนื่องจากข้อความประเภทนี้จะมีแต่ข้อมูลที่เกี่ยวกับลักษณะของการแสดงเท่านั้น, ไม่มีข้อมูลที่เกี่ยงกับโครงสร้างและความสัมพันธ์ภายในข้อความ จึงไม่เหมาะที่จะใช้ในการประมวลผล. นอกจากนั้นข้อมูลประเภทนี้ส่วนใหญ่ยังถูกออกแบบสำหรับการใช้เฉพาะของแต่ละระบบเท่านั้น.

```
<html>
<head>
<title>
ข่าวสารบนเครือข่าย
< /title>
< /head>

<body>
<h1>
ข่าวสารบนเครือข่าย
< /h1>

<p>ในปัจจุบัน, ข้อมูลข่าวสารสามารถแพร่หลายไปได้อย่างรวดเร็ว. ทั้งนี้ก็ เนื่องมาจากสาเหตุที่สำคัญสองประการคือ
<ol>
<li> ความแพร่หลายของระบบเครือข่าย ที่ทำให้บุคคลทั่วไปสามารถเข้าถึงระบบได้โดยง่าย
<li> พัฒนาการของเทคโนโลยีต่างๆ ที่ใช้บนระบบเครือข่าย ดังเช่น Web browser, ระบบ สืบค้นข้อมูลบนระบบเครือข่าย, และอื่นๆ อีกมากมาย ที่ทำให้สามารถเข้าถึงข้อมูลและข่าวสาร ได้อย่างแหม่นยำและรวดเร็ว.
< /ol>

< /body>
< /html>
```

Figure 1: ข้อความที่กำกับด้วย HTML

ข่าวสารบนเครือข่าย

ในปัจจุบัน, ข้อมูลข่าวสารสามารถแพร่หลายไปได้อย่างรวดเร็ว. ทั้งนี้ก็เนื่องมาจากสาเหตุที่สำคัญสองประการคือ

1. ความแพร่หลายของระบบเครือข่าย ที่ทำให้บุคคลทั่วไปสามารถเข้าถึงระบบได้โดยง่าย
2. พัฒนาการของเทคโนโลยีต่างๆ ที่ใช้บนระบบเครือข่าย ดังเช่น Web browser, ระบบ สืบค้นข้อมูลบนระบบเครือข่าย, และอื่นๆ อีกมากมาย ที่ทำให้สามารถเข้าถึงข้อมูลและข่าวสารได้อย่างแหม่นยำและรวดเร็ว.

Figure 2: ตัวอย่างการแสดงผลแบบที่หนึ่ง

ข่าวสารบนเครือข่าย

ในปัจจุบัน, ข้อมูลข่าวสารสามารถแพร่หลายไปได้อย่างรวดเร็ว. ทั้งนี้ก็เนื่องมาจากสาเหตุที่สำคัญสองประการคือ

1. ความแพร่หลายของระบบเครือข่าย ที่ทำให้บุคคลทั่วไปสามารถเข้าถึงระบบได้โดยง่าย
2. พัฒนาการของเทคโนโลยีต่างๆ ที่ใช้บนระบบเครือข่าย ดังเช่น Web browser, ระบบสืบค้นข้อมูลบนระบบเครือข่าย, และอื่นๆ อีกมากมาย ที่ทำให้สามารถเข้าถึงข้อมูลและข่าวสารได้อย่างแหม่นยำและรวดเร็ว.

Figure 3: ตัวอย่างการแสดงผลแบบที่สอง

3 ความเป็นจริงของภาษาที่สะท้อนจากข้อมูลจริง

กล่าวกันว่าในโลกนี้มีมากกว่า 3,500 ภาษา [5], เป็นภาษาที่ตายไปแล้วก็มีอยู่มาก. "ภาษาเป็น" เท่านั้นที่ยังมีการเปลี่ยนแปลงอยู่. ในที่นี้ผู้เขียนจะไม่กล่าวถึงทฤษฎีหรือไวยากรณ์ของภาษา, แต่จะชี้ให้เห็นถึงความสำคัญในการที่จะต้องศึกษาภาษาจากที่ใช้กันอยู่จริง.

ขณะนี้เรามีพจนานุกรมให้เลือกใช้กันมากขึ้น. กระนั้นก็ตามพจนานุกรมที่เรายึดถือใช้กันเป็นหลัก อยู่ในตอนนี้ก็คือ พจนานุกรมฉบับราชบัณฑิตยสถาน. อาจเป็นเพราะว่าเป็นพจนานุกรมที่ได้รับ การกลั่นกลองและตรวจสอบอย่างระมัดระวังมากที่สุดฉบับหนึ่ง จึงได้รับการอ้างอิงมาก, โดยเฉพาะในการตรวจสอบความถูกต้องของภาษาไทย. แต่ไม่ว่าจะเนื่องด้วยสาเหตุใดก็ตาม, จำนวนคำศัพท์ที่ปรากฏอยู่ในฉบับ พ.ศ.2525 นั้นยังมีอยู่จำกัดมาก (ประมาณ 30,000 คำ). จากที่ได้ทดลองสุ่มตรวจดูแล้ว ผู้เขียนพบว่าคำส่วนใหญ่ในพจนานุกรมฉบับราชบัณฑิตยสถาน จะเป็นคำย่อยเสียส่วนใหญ่. ส่วนใหญ่จะเป็นคำที่มีจำนวนพยางค์อยู่ระหว่าง 2-4 พยางค์ [4]. คำที่ใหญ่ขึ้น หรือคำประสมนั้น ก็มักจะเป็นคำที่มีใช้กันมานานและไม่ค่อยจะปรากฏให้เห็นในบทความที่เขียนขึ้นในปัจจุบัน. ตัวอย่างเช่น ในพจนานุกรมฉบับราชบัณฑิตยสถานมีคำว่า "ที่", "อยู่", "คุ้ม", "ค่า", "ทำ", "งาน", "ถุง", " มือ", "ตู้", และ "เย็น", แต่ไม่มีคำว่า "ที่อยู่", "คุ้มค่า", "ทำงาน", " ถุงมือ", และ "ตู้เย็น".

การกำหนดคำเพื่อที่จะบรรจุในพจนานุกรมนั้นเป็นเรื่องที่จะต้องวิจัยกันอีกมาก. การกำหนดคำอาจต้องคำนึงความถี่ของคำที่ปรากฏ ซึ่งอาจจะมองได้สองมุมที่ตรงข้ามกัน คือ: 1) ความบ่อยครั้งของการใช้สายอักขระนั้นๆ น่าจะเป็นเกณฑ์ที่จะบอกว่านั่นคือ "คำ", กับ 2) สายอักขระที่ไม่ค่อยปรากฏ มักจะยากแก่การใช้หรือเข้าใจ, ฉะนั้นควรจะบรรทึกไว้เป็น "คำ ". การจะเลือกวิธีใดนั้น ขึ้นอยู่กับจุดประสงค์ในการเตรียมพจนานุกรมมากกว่า. แต่เราก็จำเป็นที่จะต้องอ้างอิงคำที่ได้จากทั้งสองวิธี. ผู้เขียนได้เสนอการใช้วิธีการทางสถิติมาช่วยในการคัดเลือกคำตามความคิดที่หนึ่ง [6]. วิธีการนี้จะให้รายการคำที่น่าจะบรรทึกไว้ในพจนานุกรม โดยลำดับตามจำนวนครั้งที่ปรากฏในบทความนั้นๆ. ทั้งนี้ฝ่ายบัญญัติคำศัพท์จะต้อง

ทำการคัดเลือกอีกครั้ง. ส่วนวิธีการสำหรับการคัดเลือกคำตามความคิดที่สองและการทำให้วิธีการที่หนึ่งเป็นไปอย่างมีประสิทธิภาพยิ่งขึ้นนั้น ก็เป็นหัวข้อที่จะต้องทำการวิจัยต่อไป.

นอกจากคำศัพท์แล้ว, ไวยากรณ์กับการใช้ภาษาก็เป็นหัวข้อที่เราจำเป็นต้องเอาใจใส่. ผู้เขียนได้ยกปัญหาของการใช้ภาษา และสนับสนุนการใช้เครื่องหมายวรรคตอน เพื่อเป็นการแก้ปัญหาวิธีหนึ่งมาแล้ว [7]. เดิมที, ภาษาไทยนั้นยากต่อการที่จะเขียนไวยากรณ์ให้ครอบคลุมอยู่แล้ว. ปัจจุบันนี้ภาษาไทยได้เปลี่ยนแปลงไป ยิ่งทำให้ยากต่อการที่จะอธิบายลักษณะการใช้อย่างมีประสิทธิภาพได้. จะต้องมีข้อยกเว้นอยู่มากมาย. วิธีการหนึ่งที่จะทำให้ภาษาเป็นไปตามกฎเกณฑ์ที่รัดกุมได้วิธีหนึ่งก็คือการอาศัยเครื่องหมายวรรคตอน เพื่อให้ผู้เขียนยึดและคำนึงถึงการเขียนให้เป็นประโยคและรัดกุมได้. ผลที่ได้รับจากการพยายามที่จะให้ได้ภาษาที่อิงไวยากรณ์ที่ใกล้เคียงกันนี้ จะทำให้สามารถรวบรวมการใช้ภาษาได้อย่างมีประสิทธิผล. ผู้จึงได้สนับสนุนการใช้เครื่องหมายวรรคตอน และเสนอให้มีการรวบรวมคลังข้อความ. และเพื่อการแสดงให้เห็นถึงผลของการใช้ข้อมูลดังกล่าว, ผู้เขียนก็ได้แสดงผลของการใช้งานในโครงการ Emacs เพื่อภาษาไทย และการใช้งานในการสร้างระบบเพื่อการแลกเปลี่ยนข้อมูลต่างภาษาในโครงการ UNL.

4 แผนงาน ORCHID

ORCHID เป็นแผนงานเพื่อสนับสนุนการร่วมกันสร้าง, ร่วมกันใช้, และร่วมกันพัฒนาทรัพยากรทางภาษาของภาษาไทย โดยมีจุดประสงค์หลักอยู่สองประการคือ เพื่อการแก้ปัญหาเรื่องกำแพงทางภาษา และการรักษาไว้เพื่อความคงอยู่ของภาษาและวัฒนธรรมไทย.

แผนงานนี้จะรวมถึงการสร้างทรัพยากรทางภาษาโดยให้สอดคล้องตามมาตรฐาน, การศึกษาวิธีการใช้ข้อมูลทางภาษาเพื่อเป็นประโยชน์ต่อการประมวลผลภาษา, และการพัฒนาระบบเพื่อการใช้งานที่ประโยชน์ต่อสังคม. ทั้งสามประการนี้จะเป็นทั้งการสร้าง, การค้นหาวิธีการ, และการตรวจสอบจากการใช้งานจริง. แผนงานนี้จึงถูกแบ่งออกเป็นสองกลุ่ม คือ กลุ่มทรัพยากรพื้นฐาน (Basic Resources) และกลุ่มระบบประยุกต์ (Application Systems).

- Basic Resources
  - Language Processing Library and Supporting Tools
  - ORCHID POS Tagged Corpus
  - ORCHID Treebank
  - Markup Language
  - Concept Alignment
- Application Systems
  - UNL
  - Emacs for Thai
  - LEXiTRON
  - LaTeX for Thai

5 บทส่งท้าย

แผนงาน ORCHID ในบทความนี้เป็นการเริ่มต้นของผู้เขียน. บางโครงการก็กำลังดำเนินการอยู่ บางโครงการก็เป็นเพียงระยะเริ่มต้นที่ผู้เขียนได้ลงมือดำเนินการไปแล้วบ้าง และบางเรื่องก็ยังเพิ่งอยู่ในห้องทดลอง. ผู้เขียนจะได้นำแผนงานนี้เสนอต่อไปต่อองค์กรที่เกี่ยวข้อง. จุดประสงค์ของการนำเสนอครั้งนี้ก็เพื่อก่อให้เกิดความเข้าใจ, จะได้ช่วยกันคิด, ช่วยกันส่งเสริมไปในแนวทางที่พัฒนาขึ้นได้ต่อๆ ไป. สนใจหรือต้องการแนะนำ, ติดต่อผู้เขียนได้ตาม email ข้างต้น หรือดูข้อมูลเพิ่มเติมได้ที่ http://www.links.nectec.or.th/virach/home.html.

References

[1] http://www.utoronto.ca/webdocs/Official/intro.html.

[2] http://www.oasis-open.org/.

[3] http://www.personal.u-net.com/ sgml/xmlintro.htm.

[4] http://www.links.nectec.or.th/virach/research.html.

[5] Takashi, K., Rokuro, K. and Eiichi, C. 1988. The Sanseido Encyclopaedia of Linguistics.

[6] Sornlertlamvanich, V. and Hozumi, T. 1996. *The Automatic Extraction of Open Compounds from Text Corpora.* COLING-96, pp. 1143-1146.

[7] วิรัช ศรเลิศล้ำวาณิช 1998. วารสารศักยภาพ.