

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อรวบรวมข้อมูลภาษาไทยสร้างเป็นคลังข้อมูลแบบสังเคราะห์ที่มีขนาดไม่ต่ำกว่า 500 ล้านคำ และนำข้อมูลเหล่านี้มาศึกษาลักษณะการใช้ภาษาไทย โดยสนใจศึกษาในสองประเด็นหลัก คือ การพิมพ์คำหรือสะกดคำผิดอักขรวิธี และการสร้างคำใหม่ที่ยังไม่มีในพจนานุกรมมาก่อน การรวบรวมข้อมูลภาษาไทยสามารถรวบรวมได้มากกว่าที่ตั้งเป้าไว้ แต่ข้อมูลที่รวบรวมจากแหล่งต่าง ๆ มีขนาดไม่เท่ากันตามลักษณะปริมาณข้อมูลที่เกิดขึ้น การศึกษาคำสะกดผิดและคำปรากฏใหม่ในงานนี้ทำโดยใช้โปรแกรมตัดคำภาษาไทยที่มีอยู่เป็นเครื่องมือสำคัญในการกรองข้อความที่ไม่สามารถตัดคำได้ ซึ่งเหตุที่ตัดคำไม่ได้เพราะมีคำที่สะกดผิดหรือยังไม่บรรจุไว้ในพจนานุกรมที่ใช้ในการตัดคำ คำที่ไม่พบในพจนานุกรม นอกจากคำปรากฏใหม่แล้วยังมีคำประเภทอื่น ได้แก่ ชื่อเฉพาะ คำทับศัพท์ คำอุทาน คำภาษาถิ่น คำซ้ำ คำย่อ และคำพวน เมื่อแยกประเภทของคำได้แล้วจึงนำเฉพาะคำที่เป็นคำสร้างขึ้นใหม่รวมถึงคำที่แผลงมาจากคำเดิมซึ่งส่วนมากเป็นคำที่ใช้ในภาษาพูดมาวิเคราะห์

ผลการวิเคราะห์คำที่สร้างขึ้นใหม่พบว่าส่วนใหญ่เป็นคำแผลงจากคำเดิม โดยการแผลงเสียงพยัญชนะ สระ หรือวรรณยุกต์ อย่างใดอย่างหนึ่ง แต่บางครั้งก็มีการแผลงมากกว่าหนึ่งอย่าง บางกรณีก็เป็นการแผลงเฉพาะรูปสะกดโดยใช้รูปเขียนอื่นที่แทนเสียงเดียวกันในภาษาไทยได้ ส่วนคำที่สร้างขึ้นใหม่โดยไม่ได้แผลงจากคำเดิมพบได้น้อยกว่า ซึ่งผู้ฟังที่ไม่เคยรู้จักคำนั้นมาก่อนอาจมีปัญหาในการเข้าใจความหมายของคำนั้นได้ ส่วนการวิเคราะห์การสะกดคำผิด พบว่า ข้อผิดพลาดส่วนใหญ่ 55% มาจากการพิมพ์อักษรผิดเป็นตัวอื่นซึ่งอยู่ข้างเคียงโดยรอบในแป้นพิมพ์หรือเป็นการพิมพ์โดยไม่ได้ยกปุ่ม shift มากกว่าการพิมพ์อักษรเกิน พิมพ์อักษรตก หรือพิมพ์สลับตัวอักษร นอกจากนี้ยังพบข้อผิดพลาดจำนวนมากที่เป็นเรื่องวรรณยุกต์ซึ่งสะท้อนให้เห็นปัญหาความไม่เข้าใจหลักการผันวรรณยุกต์ภาษาไทย ในส่วนของการเทียบระหว่างคำผิดและคำสร้างใหม่ในภาษาที่เป็นทางการกับไม่เป็นทางการ พบว่าภาษาที่เป็นทางการมีคำผิดน้อยกว่า แต่ในส่วนของการศึกษาในเชิงการเปลี่ยนแปลงตามเวลายังไม่ได้ทำ เนื่องจากข้อมูลที่เก็บยังมีช่วงเวลาสั้นอีกทั้งยังไม่ได้นำข้อมูลทั้งหมดเข้าในระบบฐานข้อมูล