



การศึกษาสถานภาพงานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย ระหว่างปี 2542-2563

A Study of the Status of corpus-based researches in Thailand during 1999-2020

ช้วยหยุน เซิน¹, ทินวัฒน์ สร้อยกุดเรือ^{2*}
Cuiyun Shen, Thinnawat Sroikudrua

บทคัดย่อ

ในปัจจุบัน เทคโนโลยีถูกนำมาใช้ประโยชน์ในการศึกษาภาษาอย่างกว้างขวาง คลังข้อมูลภาษาขนาดใหญ่กลายเป็นทรัพยากรที่สำคัญในการศึกษาภาษาต่าง ๆ บทความนี้ได้รวบรวมและวิเคราะห์เนื้อหาสำคัญของงานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย ระหว่างปี 2542-2563 จำนวน 52 เรื่อง โดยมีวัตถุประสงค์เพื่อศึกษาสถานภาพงานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย เพื่อเป็นเอกสารหรือข้อมูลอ้างอิงสำหรับผู้สนใจคลังข้อมูลภาษาในประเทศไทย หรือเพื่อเป็นแนวทางในการประยุกต์ใช้คลังข้อมูลภาษามาทำการวิจัย ผลการศึกษาพบว่าเนื้อหาสำคัญของการวิจัยมีอยู่ 6 ประเภท ได้แก่ งานวิจัยด้านคำศัพท์ อรรถศาสตร์ และไวยากรณ์ งานวิจัยด้านการเรียนการสอน งานวิจัยด้านการเปรียบเทียบภาษา งานวิจัยด้านการประมวลศัพท์ งานวิจัยด้านคลังข้อมูลภาษาผู้เรียน และงานวิจัยด้านการแปลภาษาด้วยเครื่อง ส่วนสถานภาพของงานวิจัยพบว่าเนื้อหาของงานวิจัย และภาษาของคลังข้อมูลที่ใช้ในงานวิจัยขาดความสมดุลและความหลากหลาย

คำสำคัญ: งานวิจัยแบบอิงคลังข้อมูลภาษา สถานภาพงานวิจัย คลังข้อมูลภาษาของประเทศไทย

¹ นักศึกษาหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาภาษาไทย คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยขอนแก่น

PhD student in Philosophy Thai, Faculty of Humanities and Social Sciences, Khon Kaen University.

cuiyun.s@kkumail.com

² ผู้ประสานงานหลัก: ผู้ช่วยศาสตราจารย์ สาขาวิชาภาษาไทย คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยขอนแก่น Assistant

Professor of Thai Language, Faculty of Humanities and Social Sciences, Khon Kaen University. thinsr@kku.ac.th



Abstract

Nowadays, technology is widely used in language researches. Corpus has become an important resource in language research. This article aims to review and analyze the main content of researches on corpus linguistics in Thailand during the 1999-2020, amounting to 52 issues, to study the status of researches in corpus linguistic of Thailand. It is expected that this article will serve as a reference for researchers in Thai corpus linguistics. Alternatively, it may provide insights for the application of corpus in research. The investigation result found that the types of the researches include vocabulary, semantics, and grammar research; teaching and learning research; comparative linguistics research; terminology research; learner corpus research; and machine-based translation research. Moreover, the state of researches found that the fields of corpus-based researches in Thailand, and the languages of the corpus are imbalanced and lack abundance.

Key words: Corpus-based Researches, Researches Status Study, Language Corpus of Thai



บทนำ

ในปัจจุบัน เทคโนโลยีถูกนำมาใช้ประโยชน์ในการศึกษาภาษาอย่างกว้างขวาง คลังข้อมูลภาษาขนาดใหญ่กลายเป็นทรัพยากรที่สำคัญในการวิจัยภาษาต่างๆ ดังจะเห็นได้จากการตื่นตัวของนานาชาติในการจัดสร้างคลังข้อมูลภาษาขนาดใหญ่ เช่น British National Corpus (100 ล้านคำ แล้วเสร็จในปี ค.ศ. 1994), The Bank of English (ปัจจุบันมีมากกว่า 450 ล้านคำ), American National Corpus (เป้าหมาย 100 ล้านคำ ปัจจุบันเก็บได้ 22 ล้านคำ), Czech National Corpus, Hellenic National Corpus (คลังข้อมูลภาษากรีก ปัจจุบันมีมากกว่า 34 ล้านคำ), National Corpus of Irish (30 ล้านคำ แล้วเสร็จในปี ค.ศ. 1999) คลังข้อมูลเหล่านี้ถูกนำมาใช้ในงานต่าง ๆ อาทิ การจัดทำพจนานุกรม การจัดทำหนังสืออ้างอิงต่าง ๆ และใช้ในการศึกษาวิจัยทางภาษาศาสตร์อีกมากมาย (วิโรจน์ อรุณมานะกุล, 2553) ประเทศไทยเริ่มมีการวิจัยด้านคลังข้อมูลภาษาในประมาณปี 2542 โดยจุฬาลงกรณ์มหาวิทยาลัยถือว่าเป็นผู้เบิกทางคลังข้อมูลภาษาของประเทศไทย

ในช่วงยี่สิบปีที่ผ่านมา การศึกษาแบบอิงคลังข้อมูลภาษาได้กลายเป็นวิธีการวิจัยที่สำคัญของภาษาศาสตร์ การวิจัยที่ประยุกต์ใช้คลังข้อมูลมีจำนวนเพิ่มขึ้นเรื่อย ๆ และกลายเป็นประเด็นที่น่าสนใจ เพื่อให้เข้าใจสถานภาพงานวิจัยแบบอิงคลังข้อมูลภาษาภายในประเทศไทยมากขึ้น บทความนี้จะดำเนินการรวบรวมและวิเคราะห์งานวิจัยแบบอิงคลังข้อมูลภาษาภายในประเทศไทยอย่างเป็นระบบ หวังว่าจะเป็นเอกสารหรือข้อมูลอ้างอิงสำหรับผู้สนใจงานวิจัยแบบอิงคลังข้อมูลภาษาในประเทศไทย หรือเพื่อเป็นแนวทางในการประยุกต์ใช้คลังข้อมูลภาษามาทำวิจัย

บทความนี้ได้เก็บรวบรวมและศึกษางานวิจัยแบบอิงคลังข้อมูลภาษาในประเทศไทยระหว่างปี 2542-2563 ทั้งหมด 52 เรื่อง เพื่อศึกษาสถานภาพงานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย โดยเนื้อหาสำคัญแบบออกเป็น 2 ส่วน ได้แก่ เนื้อหาสำคัญของงานวิจัย และสถานภาพงานวิจัย

ขอบเขตของการศึกษา

การวิจัยครั้งนี้ศึกษาเฉพาะงานวิจัยด้านภาษาศาสตร์แบบอิงคลังข้อมูลภาษาของประเทศไทยระหว่างปี 2542-2563 ที่ตีพิมพ์หรือเผยแพร่เป็นภาษาไทยเท่านั้น รวมงานวิจัยที่ตีพิมพ์ในวารสารวิจัยที่เป็นวิทยานิพนธ์ ดุษฎีนิพนธ์ และงานวิจัยฉบับสมบูรณ์ ไม่รวมงานวิจัยด้านสาขาอื่น อย่างเช่น การวิจัยด้านวิศวกรรม การวิจัยด้านภาษาศาสตร์คอมพิวเตอร์



วิธีดำเนินการศึกษา

บทความนี้เป็นการสังเคราะห์งานวิจัยโดยการวิเคราะห์เนื้อหา(content analysis) มีขั้นตอนการดำเนินการดังต่อไปนี้

1. สืบค้นและรวบรวมงานวิจัยทางด้านภาษาที่เกี่ยวกับคลังข้อมูลภาษาของประเทศไทยจากแหล่งข้อมูลต่าง ๆ ได้แก่
 - 1) The Chulalongkorn University Intellectual Repository (CUIR) คลังปัญญาจุฬาฯ
 - 2) Thammasat University Digital Collections. คลังเอกสารออนไลน์ ในคลังของธรรมศาสตร์
 - 3) Dspace ฐานข้อมูลมหาวิทยาลัยศิลปากร
 - 4) PSU Knowledge Bank, คลังความรู้มหาวิทยาลัยสงขลานครินทร์
 - 5) ThaiLIS.
 - 6) Digital Research Information Center. ฐานข้อมูลที่เผยแพร่งานวิจัยและวิทยานิพนธ์จากนักวิจัยและมหาวิทยาลัยต่าง ๆ
 - 7) และ <https://scholar.google.com/>
2. วิเคราะห์เนื้อหาสำคัญของงานวิจัย
3. จัดทำตารางบันทึกและเปรียบเทียบข้อมูลสำคัญของการวิจัย
4. วิเคราะห์สภาพรวมทั้งปัญหา
5. เขียนบรรยายผลการศึกษา

ผลการศึกษา

1. ผลการศึกษาด้านเนื้อหาสำคัญของงานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย

จากการรวบรวมและวิเคราะห์งานวิจัยแบบอิงคลังข้อมูลภาษาในประเทศไทยระหว่างปี 2542-2563 ทั้งหมด 52 เรื่อง สามารถสรุปจำนวนตามเวลา ประเภทและเนื้อหาสำคัญของงานวิจัยดังตารางต่อไปนี้



ตารางที่ 1 ตารางแสดงจำนวนงานวิจัยตามเวลา



ตารางที่ 2 ตารางแสดงประเภทของงานวิจัย

ประเภทงานวิจัย	วิทยานิพนธ์ปริญญาโท	บทความวิจัย	งานวิจัยฉบับสมบูรณ์	ดุสิตนิพนธ์
จำนวน	34	10	4	4
ร้อยละ	65.38	19.22	7.70	7.70

ตารางที่ 3 ตารางแสดงเนื้อหาสำคัญของงานวิจัย

เนื้อหาสำคัญ	คำศัพท์ อรรถศาสตร์ และ ไวยากรณ์	การเรียนรู้ การสอน	การ เปรียบเทียบ ภาษา	ประมวล ศัพท์	คลังข้อมูล ภาษาผู้เรียน	การ แปลภาษา ด้วยเครื่อง
จำนวน	26	11	6	5	2	2
ร้อยละ	50	21.15	11.54	9.61	3.85	3.85

จากตารางที่ 1 สะท้อนให้เห็นว่างานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทย ระหว่างปี 2542-2563 มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ แสดงให้เห็นว่ามีผู้วิจัยที่เห็นถึงประโยชน์และความสำคัญของคลังข้อมูลภาษาเพิ่มขึ้น



จากตารางที่ 2 จะเห็นได้ว่า ประเภทของงานวิจัยมี 4 ประเภท ดังต่อไปนี้ วิทยานิพนธ์ปริญญาโทมีจำนวนมากที่สุด 34 เรื่อง คิดเป็นร้อยละ 65.38 บทความวิจัยมีจำนวน 10 เรื่อง คิดเป็นร้อยละ 19.22 งานวิจัยฉบับสมบูรณ์จำนวน 4 เรื่อง คิดเป็นร้อยละ 7.70 ดุษฎีนิพนธ์จำนวน 4 เรื่อง คิดเป็นร้อยละ 7.70

จากตารางที่ 3 จะเห็นได้ว่าเนื้อหาสำคัญของการวิจัยด้านภาษาศาสตร์คลังข้อมูลภาษาของประเทศไทยมี 6 ประเภท ดังต่อไปนี้ งานวิจัยด้านคำศัพท์ วรรคศาสตร์ และไวยากรณ์ มีจำนวนมากที่สุด 26 เรื่อง คิดเป็นร้อยละ 50 งานวิจัยด้านการเรียนการสอนมีจำนวน 11 เรื่อง คิดเป็นร้อยละ 21.15 งานวิจัยด้านการเปรียบเทียบภาษามีจำนวน 6 เรื่อง คิดเป็นร้อยละ 11.54 งานวิจัยด้านการประมวลผลมีจำนวน 5 เรื่อง คิดเป็นร้อยละ 9.61 งานวิจัยวิจัยด้านคลังข้อมูลภาษาผู้เรียนมีจำนวน 2 เรื่อง คิดเป็นร้อยละ 3.85 งานวิจัยด้านการแปลภาษาด้วยเครื่องมีจำนวน 2 เรื่อง คิดเป็นร้อยละ 3.85 แต่ละประเภทมีรายละเอียดดังต่อไปนี้

1.1 งานวิจัยด้านการศึกษาคำศัพท์ วรรคศาสตร์ และไวยากรณ์

แต่เดิมการศึกษาคำศัพท์เน้นศึกษาความหมายต่าง ๆ ของคำศัพท์ และความแตกต่างกันระหว่างคำพ้องความหมายเป็นหลัก การพัฒนาของคลังข้อมูลภาษาทำให้ขอบเขตของการศึกษาคำศัพท์ได้รับการขยายมากขึ้น จากการศึกษาพบว่างานวิจัยด้านคำศัพท์ วรรคศาสตร์ และไวยากรณ์เป็นประเภทงานวิจัยที่มีจำนวนมากที่สุดที่ เนื่องจากการใช้คลังข้อมูลภาษาร่วมกับโปรแกรมคลังข้อมูลภาษาช่วยให้ผู้วิจัยเข้าใจความหมายและการใช้คำ ๆ นั้นได้ดีขึ้น ได้เห็นตัวอย่างและความหลากหลายของการใช้คำ ๆ นั้นและช่วยให้ผู้ศึกษาสามารถเข้าใจความหมายต่าง ๆ ของคำนั้นได้ดีและถูกต้องยิ่งขึ้น งานวิจัยประเภทนี้สามารถยกตัวอย่างได้ดังต่อไปนี้

ชาตบุษย์ นิสระรัตน์ (2551) ทำการวิจัยเรื่อง “คำปรากฏร่วมจำเพาะและอรรถสัมผัสของคำกริยา “เกิด” ในภาษาไทย-การศึกษาเชิงภาษาศาสตร์คลังข้อมูล” โดยมีวัตถุประสงค์เพื่อวิเคราะห์คำที่ปรากฏร่วมจำเพาะและอรรถสัมผัสของคำกริยา “เกิด” และเพื่อเปรียบเทียบการปรากฏร่วมจำเพาะและอรรถสัมผัสของคำกริยา “เกิด” ระหว่างข้อมูลต่างประเภท

พิรุฬห์ ปิยมพงศ์ (2559) ทำการวิจัยเรื่อง “คุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์ของตัวบ่งชี้ผู้รับประโยชน์ “เพื่อ” และ “ให้” ในภาษาไทย” โดยการวิจัยนี้ได้เก็บและคัดเลือกประโยคที่มีคำว่า “เพื่อ” และ “ให้” จากคลังข้อมูลภาษาไทยแห่งชาติ วิเคราะห์และเปรียบเทียบคุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์ของตัวบ่งชี้ผู้รับประโยชน์ “เพื่อ” และ “ให้” ในภาษาไทย และศึกษาปัจจัยที่มีผลต่อคุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์ที่แตกต่างกัน



ปิยะวดี คำสุวรรณ, คเชนทร์ ตัญศิริ, อุมารณ สักขมาน (2562) ทำการวิจัยเรื่อง “การศึกษาคำว่า “ตาม” ในภาษาไทย” เพื่อศึกษาวิเคราะห์คุณสมบัติทางความหมายของคำว่า “ตาม” ในภาษาไทยปัจจุบันจากมุมมองภาษาศาสตร์ปริชาน โดยวิเคราะห์ตัวอย่างข้อมูลภาษา จำนวนทั้งหมด 1,000 ข้อมูล จากคลังข้อมูลภาษาไทยแห่งชาติ ฯ

จิตสุตา บุญรัตน์ , คเชนทร์ ตัญศิริ, มุกข์ดา สุขธราจาร (2563) ทำการศึกษาคุณสมบัติทางความหมายของการปรากฏร่วมของคำว่า “ติด” ในภาษาไทยโดยใช้คลังข้อมูลภาษา เพื่อศึกษารูปแบบการปรากฏร่วมจำเพาะ(Collocation) ของคำกริยา “ติด” ในภาษาไทยโดยวิเคราะห์จากคลังข้อมูลภาษาและวิเคราะห์คุณสมบัติทางความหมายของคำกริยา “ติด” จากบริบทของการปรากฏร่วม

ผไทมาศ ประมวล , พุทธชาติ โพธิบาล , คเชนทร์ ตัญศิริ (2563) ทำการวิจัยเรื่อง “เครือข่ายความหมาย ‘นอก’ ในภาษาไทย” งานวิจัยนี้มุ่งศึกษาคุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์เพื่อหาเครือข่ายความหมาย ของคำว่า นอก ตามแนวคิดทางอรรถศาสตร์ปริชาน โดยวิเคราะห์จากคำ วลี พร้อมทั้งปรับบริบทจากคลังข้อมูลภาษาไทยแห่งชาติ จำนวน 2,000 ตัวอย่าง

นพวรรณ เมืองแก้ว (2556) ได้ศึกษาคำแสดงการขอร้องในภาษาไทย เพื่อศึกษาที่มาของคำแสดงการขอร้องในภาษาไทยจำนวน 9 คำ ได้แก่ คำว่า “กรุณา” “ขอ” “ช่วย” “เชิญ” “โปรด” “รบกวน” “ด้วย” “ที” และ “หน่อย” ในแง่ปัจจัยและกระบวนการที่ทำให้คำเหล่านี้กลายมาเป็นคำแสดงการขอร้องในภาษาไทย งานวิจัยนี้เป็นการศึกษาตามแนวทฤษฎีการกลายเป็นคำไวยากรณ์ ข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลเฉพาะภาษาไทยกรุงเทพฯ ปัจจุบัน โดยเก็บข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติ

จะเห็นได้ว่างานวิจัยประเภทนี้ เน้นการศึกษาระดับคำ ไม่ว่าจะเป็นคุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์ของคำ คุณสมบัติทางความหมายของคำ คำปรากฏร่วมจำเพาะและอรรถสัมพันธ์ของคำ หรือเครือข่ายความหมายของคำ

1.2 งานวิจัยด้านการเรียนการสอนภาษา

เนื่องจากคลังข้อมูลเป็นแหล่งให้ข้อมูลการใช้ภาษาที่เกิดขึ้นจริง คลังข้อมูลภาษาจึงเป็นแหล่งข้อมูลที่สำคัญต่อการศึกษาวิจัยทางด้านภาษาและการสอนภาษา การวิจัยด้านการเรียนการสอนเป็นการวิจัยที่มีจำนวนอยู่ที่ลำดับ 2 จำนวน 11 เรื่อง โดยสาขาวิชาที่สังกัดมีสาขาวิชาภาษาศาสตร์ สาขาวิชาภาษาไทย สาขาวิชาภาษาอังกฤษ สาขาวิชาการสอนภาษาจีน สาขาวิชาภาษาศาสตร์ประยุกต์ และสาขาวิชาการสอนภาษาอังกฤษ สามารถยกตัวอย่างงานวิจัยได้ดังต่อไปนี้



Zhao Jiaoyue (2558) ทำการวิจัยเรื่อง “การศึกษาความแตกต่างการใช้คำศัพท์ และวิธีการสอนภาษาจีนเกี่ยวกับปลา งู หนอน เต่า ในคลังคำศัพท์” โดยผู้วิจัยเก็บข้อมูลจากคลังข้อมูลภาษาจีนโดยใช้คำภาษาจีน 4 คำนี้ วิเคราะห์ความหมาย และได้นำเสนอกลยุทธ์การสอนคำภาษาจีน 4 คำนี้

ธารทอง แจ่มไพบูลย์ (2559) ทำการวิจัยเรื่อง “พัฒนาการของคำศัพท์ที่มีความถี่ในการปรากฏสูงในหนังสือเรียนวิชาภาษาไทย: การศึกษาตามแนวทางภาษาศาสตร์คลังข้อมูล” โดยมีวัตถุประสงค์เพื่อศึกษาพัฒนาการของคำศัพท์ที่มีความถี่ในการปรากฏสูงในหนังสือเรียนวิชาภาษาไทย โดยมีการเปรียบเทียบการขยายวงศัพท์ รวมถึงการขยายขอบเขตทางความหมายและหน้าที่ในสองมิติคือเปรียบเทียบระหว่างช่วงชั้น และเปรียบเทียบระหว่างหลักสูตร โดยมีสมมติฐานว่าช่วงชั้นที่สูงขึ้นจะมีวงศัพท์มากกว่าและครอบคลุมช่วงชั้นที่ต่ำกว่า นอกจากนี้ยังมีความหมายและหน้าที่ของคำศัพท์เพิ่มมากขึ้นตามช่วงชั้นด้วย ส่วนในเชิงหลักสูตรมีสมมติฐานว่า แต่ละหลักสูตรมีวงศัพท์ร่วมกันร้อยละ 90 ขึ้นไป รวมทั้งมีความหมายและหน้าที่เหมือนกันทุกหลักสูตร งานวิจัยชิ้นนี้ศึกษาตามแนวทางภาษาศาสตร์คลังข้อมูล จากคลังข้อมูลหนังสือเรียนวิชาภาษาไทย (TTC) ขนาด 3,037,772 คำที่สร้างขึ้น ซึ่งมีจำนวนศัพท์รวมทั้งสิ้น 19,494 รูปศัพท์

ปิยะพงษ์ เลาศรีรัตนชัย (2562) ทำการวิจัยเรื่อง “การสร้างรายการคำศัพท์เพื่อธุรกิจบริการและบทเรียนคำศัพท์ออนไลน์” โดยเก็บรวบรวมข้อมูลเก็บข้อมูลเกี่ยวกับธุรกิจบริการจากนิตยสารการท่องเที่ยว 2560-2561 คู่มือปฏิบัติการงานสายการบินและเว็บไซต์เกี่ยวกับแวดวงบริการจำนวน 300 เว็บไซต์ และสร้างเป็นคลังข้อมูลภาษาขนาด 52,079,960 คำ และใช้รายการคำศัพท์ความถี่สูงที่ได้มาจากคลังข้อมูลภาษามาสร้างเป็นบทเรียนออนไลน์ ทดลองใช้กับนักศึกษากลุ่มตัวอย่างเพื่อพัฒนาความสามารถในด้านการสื่อสารเชิงธุรกิจบริการของนักศึกษาคนไทย

จากงานวิจัยที่กล่าวมา จะเห็นได้ว่าเนื้อหาสำคัญของงานวิจัยเหล่านี้ มีการใช้คลังข้อมูลภาษามาช่วยในการเข้าใจความหมายและบริบทการใช้ และการใช้คลังข้อมูลมาช่วยในการสร้างรายการคำศัพท์ และบทเรียน

1.3 งานวิจัยด้านการเปรียบเทียบภาษา

การเปรียบเทียบภาษาแบบอิงคลังข้อมูลภาษา ถือว่าเป็นงานวิจัยประเภทหนึ่งที่ได้รับประโยชน์จากคลังข้อมูลภาษาเป็นอย่างมาก การเปรียบเทียบภาษาแบบอิงคลังข้อมูลภาษาโดยทั่วไปแล้วเป็นการเปรียบเทียบภาษาระดับคำ มีด้านความหมายของคำ อุปลักษณะของคำ มโนทัศน์ของคำ ความหมายเชิงเปรียบเทียบของคำ เป็นต้น เป็นงานวิจัยที่ได้ใช้ประโยชน์จาก



การศึกษาคำปรากฏร่วม (collocations) และความถี่ของการปรากฏ (frequency) จากคลังข้อมูลภาษา จากการศึกษาพบจำนวนงานวิจัยด้านนี้ทั้งหมด 6 เรื่อง โดยมี 3 เรื่องสังกัดสาขาวิชาภาษาศาสตร์ 2 เรื่องสังกัดสาขาวิชาการสอนภาษาจีน และอีก 1 เรื่องสังกัดสาขาภาษาศาสตร์เพื่อการสื่อสาร สามารถยกตัวอย่างงานวิจัยได้ดังต่อไปนี้

รังสิมา รุ่งเรือง (2560) ศึกษาเรื่อง “ความหมายเปรียบเทียบของสัตว์ในภาษาไทยและภาษาอินโดนีเซีย” โดยเก็บรวบรวมข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติ (Thai National Corpus) และคลังข้อมูลภาษาอินโดนีเซีย ห้องสมุดภาษาเอเชียตะวันออกเฉียงใต้ (SEALang Library Indonesian Corpus) เพื่อศึกษาและวิเคราะห์ความหมายเปรียบเทียบตลอดจนกระบวนการทางการเปรียบเทียบสัตว์ในภาษาไทยและภาษาอินโดนีเซีย รวมถึงเพื่อศึกษาและอธิบายโลกทัศน์ ค่านิยม และทัศนคติของชาวไทยและชาวอินโดนีเซียจากความหมายเปรียบเทียบของสัตว์ในภาษาไทยและภาษาอินโดนีเซีย

ชื่อนี เหยา (2561) ทำการศึกษาระบบมโนทัศน์ “ใจ” ในภาษาไทยเปรียบเทียบกับภาษาจีน : การศึกษาตามแนวภาษาศาสตร์ปริชาน โดยเก็บรวบรวมข้อมูลประโยคภาษาไทยจากคลังข้อมูลภาษาไทยแห่งชาติ และเก็บรวบรวมข้อมูลประโยคภาษาไทยคลังข้อมูลภาษาจีนสมัยใหม่ของมหาวิทยาลัยภาษาและวัฒนธรรมแห่งกรุงปักกิ่ง เพื่อศึกษาและวิเคราะห์อุปลักษณ์ “ใจ” ในภาษาไทย รวมถึงอธิบายถึงมโนทัศน์ “ใจ” ในภาษาไทยที่สะท้อนจากการใช้คำว่า “ใจ” เป็นอุปลักษณ์ ตลอดจนเพื่อเปรียบเทียบมโนทัศน์ “ใจ” ในภาษาไทยกับมโนทัศน์ ใจ/心 (jīn55) ในภาษาจีนตามแนวคิดทฤษฎีด้านภาษาศาสตร์ปริชาน

อภิชาติ คานนิม (2561) ทำการศึกษามโนทัศน์การเคลื่อนที่เข้าในภาษาไทยและภาษาญี่ปุ่น โดยเก็บข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติและคลังข้อมูลภาษาญี่ปุ่น (Balanced Corpus of Contemporary Written Japanese) เพื่อศึกษาคำกริยาแสดงการเคลื่อนที่เข้าในภาษาไทยและภาษาญี่ปุ่นและเปรียบเทียบมโนทัศน์การเคลื่อนที่เข้าในภาษาไทยและภาษาญี่ปุ่น

จากที่กล่าวมา จะเห็นได้ว่างานวิจัยการเปรียบเทียบภาษาของไทยเน้นการเปรียบเทียบภาษาระดับคำ รวมเปรียบเทียบอุปลักษณ์ของคำ มโนทัศน์ของคำ โลกทัศน์ ค่านิยม ทัศนคติที่สะท้อนผ่านการใช้คำ และเน้นการเปรียบเทียบคำในภาษาไทยกับคำในภาษาต่างประเทศ และจากการศึกษาครั้งนี้ยังพบว่า คลังข้อมูลภาษาไทยที่ใช้ในการวิจัยด้านนี้ล้วนเป็นคลังข้อมูลภาษาไทยแห่งชาติของจุฬาลงกรณ์มหาวิทยาลัย



1.4 งานวิจัยด้านการประมวลศัพท์

การประมวลศัพท์ (Terminology) คล้ายกับการทำพจนานุกรม (Lexicography) สิ่งที่แตกต่างกันคือขอบเขตเนื้อหาพจนานุกรมจะครอบคลุมภาษาโดยรวมในขณะที่ประมวลศัพท์จะครอบคลุมศัพท์เฉพาะสาขา (วิโรจน์ อรุณมานะกุล, 2553) จากการศึกษาพบจำนวนงานวิจัยประเภทนี้ทั้งหมด 5 เรื่อง ล้วนสังกัดสาขาวิชาการแปลและการล่าม โดยมีงานวิจัย 4 เรื่องสังกัดจุฬาลงกรณ์มหาวิทยาลัย และอีก 1 เรื่องสังกัดมหาวิทยาลัยธรรมศาสตร์ และจากการวิเคราะห์เนื้อหาของงานวิจัยทั้ง 5 เรื่องนี้ พบว่าคลังข้อมูลภาษาที่ถูกสร้างขึ้น ล้วนเป็นคลังข้อมูลภาษาอังกฤษ เมื่อได้รายการคำศัพท์แล้ว ค่อยดึงออกมา แล้ววิเคราะห์เลือกคำศัพท์ภาษาไทยที่นำมาสร้างมโนทัศน์ โดยสามารถยกตัวอย่างงานวิจัยได้ดังต่อไปนี้

ชุตินา เตชะพิรุณห์ (2550) ทำการวิจัยเรื่อง “ประมวลศัพท์เรื่องเครื่องหมายการค้า” ศึกษาระเบียบวิธีการประมวลศัพท์ตามทฤษฎีทางศัพท์วิทยา และจัดทำประมวลศัพท์เรื่องเครื่องหมายการค้า ซึ่งประกอบด้วยศัพท์เกี่ยวกับการจดทะเบียนเครื่องหมายการค้า โดยผู้วิจัยเก็บรวบรวมข้อมูลที่เกี่ยวข้องและสร้างเป็นคลังข้อมูลภาษาอังกฤษ แล้วดึงคำออกมาจากคลังข้อมูลภาษา แล้ววิเคราะห์เลือกคำศัพท์ภาษาไทยที่นำมาสร้างมโนทัศน์ เสร็จแล้วสร้างมโนทัศน์สัมพันธ์และบันทึกข้อมูลศัพท์ ประมวลศัพท์งานวิจัยนี้มุ่งให้เป็นเอกสารอ้างอิงสำหรับนักแปลและผู้สนใจทั่วไป

ปาริสสา เหล่าวนิชย์ (2552) ทำการวิจัยเรื่อง “ประมวลศัพท์เรื่องการสำรวจและผลิตรายการขนส่งและรูปแบบการใช้เชื้อเพลิงก๊าซธรรมชาติ” งานวิจัยนี้คล้ายกับการวิจัยของนางสาวชุตินา เตชะพิรุณห์ (2550) โดยมีวัตถุประสงค์เพื่อศึกษากระบวนการสร้างประมวลศัพท์ และนำความรู้ที่ได้มาจัดเป็นประมวลศัพท์เรื่องการสำรวจและผลิตรายการขนส่งและรูปแบบการใช้เชื้อเพลิงก๊าซธรรมชาติ

ขวัญชนิตร์ สุทธิสารสุนทร (2546) ทำการวิจัยเรื่อง “กระบวนการทำประมวลศัพท์หมวดพาณิชย์อิเล็กทรอนิกส์” โดยมีวัตถุประสงค์เพื่อศึกษากระบวนการที่เหมาะสมในการจัดทำประมวลศัพท์ตามหลักศัพท์วิทยา เพื่อหากระบวนการที่เหมาะสมในการจัดทำประมวลศัพท์หมวดพาณิชย์อิเล็กทรอนิกส์และเพื่อจัดทำประมวลศัพท์หมวดพาณิชย์อิเล็กทรอนิกส์ และเพื่อเป็นประโยชน์ในฐานะเอกสารอ้างอิงสำหรับนักแปลและผู้สนใจพาณิชย์อิเล็กทรอนิกส์

จากงานวิจัยที่กล่าวมาทั้ง 3 เรื่องจะเห็นได้ว่างานวิจัยกลุ่มนี้มุ่งเป็นประโยชน์เพื่อใช้เป็นเอกสารอ้างอิงสำหรับนักแปลและผู้สนใจทั่วไป



1.5 งานวิจัยด้านการสร้างคลังข้อมูลภาษาผู้เรียน

คลังข้อมูลภาษาผู้เรียน (Learner Corpus) เป็นคลังข้อมูลภาษาประเภทหนึ่งที่เก็บข้อมูลภาษาที่ได้จากผู้เรียนภาษาต่างประเทศ อาจเป็นภาษาเขียนหรือเป็นภาษาพูดก็ได้ เนื่องจากเป็นข้อมูลจากผู้ที่ไม่ได้เป็นเจ้าของภาษาจึงมีลักษณะแตกต่างจากคลังข้อมูลภาษาปกติที่มักเก็บข้อมูลภาษาจากเจ้าของภาษาในหลายลักษณะ และคลังข้อมูลภาษาผู้เรียนก็เป็นข้อมูลจริงที่ทำให้เข้าใจสภาพการใช้ภาษาที่ผู้เรียนภาษานั้นตลอดจนปัญหาที่เป็นอยู่ว่ามีความเบี่ยงเบนจากภาษามาตรฐานหรือภาษาที่เจ้าของภาษาใช้กันอย่างไร อะไรที่ใช้แล้วดูไม่เป็นธรรมชาติเหมือนเจ้าของภาษา ในขณะเดียวกันก็ทำให้เห็นภาษากลาง (Interlanguage) ที่เกิดขึ้น ในระหว่างพัฒนาการเรียนรู้อาษาต่างประเทศและภาษาที่สอง คลังข้อมูลภาษาผู้เรียนจึงเป็นทรัพยากรที่เป็นประโยชน์อย่างยิ่งต่อการเรียนการสอนภาษาต่างประเทศและภาษาที่สองในด้านต่าง ๆ (วิโรจน์ อรุณมานะกุล, 2553) จากการศึกษาพบงานวิจัยที่สร้างคลังข้อมูลภาษาผู้เรียนทั้งหมด 2 เรื่อง ได้แก่ งานของกระรัตเพชร คงรอด (2560) ที่สังกัดสาขาวิชาการสอนภาษาอังกฤษ และงานของจุฬารัตน์ คำน้อย, หลิวชู, จิ้นหลิง (2561) ที่สังกัดสำนักจิ้นวิทยา มหาวิทยาลัยแม่ฟ้าหลวง 2 เรื่องนี้มีรายละเอียดดังต่อไปนี้

กระรัตเพชร คงรอด (2560) ทำการวิจัยเรื่อง “การวิเคราะห์ข้อผิดพลาดด้านการใช้คำในการเขียนภาษาอังกฤษเชิงวิชาการของนักศึกษาไทยระดับปริญญาตรี ” โดยมีวัตถุประสงค์ที่ศึกษาการใช้ข้อผิดพลาดด้านการใช้คำที่ใช้ในงานเขียนภาษาอังกฤษเชิงวิชาการของนักศึกษา ความสัมพันธ์ระหว่างข้อผิดพลาดกับสาขาวิชาของผู้เรียน และแนวทางแนวทางในการพัฒนาการเรียนการสอน การเขียนในด้านการใช้คำในรายวิชานั้น โดยผู้วิจัยได้สร้างคลังข้อมูลภาษาของนักศึกษาไทยโดยใช้ข้อมูลที่เก็บรวบรวมจากงานเขียนเชิงวิชาการของนักศึกษา 52 คน จำนวน 2 หัวข้อ (ขนาดข้อมูลภาษาที่สร้างขึ้นเป็นจำนวน 27,211 คำ) และทำตารางบันทึกและวิเคราะห์ข้อผิดพลาดด้านการใช้คำของนักศึกษา

จุฬารัตน์ คำน้อย, หลิวชู, จิ้นหลิง (2561) ทำการวิจัยเรื่อง “การสร้างคลังข้อมูลและวิเคราะห์การเขียนตัวอักษรจีนที่ผิดของนักศึกษาไทย ระดับอุดมศึกษา กรณีศึกษาสำนักวิชาจิ้นวิทยา มหาวิทยาลัยแม่ฟ้าหลวง ” โดยการวิจัยนี้ได้สร้างคลังข้อมูลการเขียนตัวอักษรจีนที่ผิดโดยใช้ข้อมูลจากการเขียนตามคำบอก การบ้าน บทความ และข้อสอบของนักศึกษาชั้นปีที่ 1-4 ของมหาวิทยาลัยแม่ฟ้าหลวง และทำการวิเคราะห์ประเภทการเขียนตัวอักษรผิดของนักศึกษาไทย และพบว่าการสร้างคลังข้อมูลดังกล่าวขึ้น สามารถนำมาแสดงผลเกี่ยวกับการเขียนตัวอักษรจีนผิดได้ ไม่ว่าจะเป็นตัวอักษรจีนที่ผู้เรียนเขียนผิด หรือลักษณะการเขียนผิดของตัวอักษรนั้น ในขณะเดียวกัน



อาจารย์ผู้สอนก็สามารถใช้คลังข้อมูลดังกล่าวเพื่อศึกษาสาเหตุการเขียนตัวอักษรจีนผิดของผู้เรียน และสามารถนำไปประกอบการสอนเพื่อป้องกันการเขียนผิดของนักศึกษาได้ นอกจากนี้ผลการศึกษายังพบว่าการเขียนตัวอักษรจีนผิดของนักศึกษาไทยนั้น มีอยู่ 6 ประเภท ได้แก่ ประเภทขีดหรือส่วนประกอบขาดหายไป ประเภทการเขียนขีดหรือส่วนประกอบข้างเกิน ประเภทการนำขีดหรือส่วนประกอบของตัวอักษรอื่นมาแทนที่ ประเภทการเขียนขีดหรือส่วนประกอบของตัวอักษรจีนผิดตำแหน่ง ประเภทการเขียนขีดหรือส่วนประกอบของตัวอักษรผิดรูป ประเภทการเขียนส่วนประกอบของตัวอักษรแบบปะปนกัน

จากงานวิจัยทั้ง 2 เรื่อง จะเห็นได้ว่าการวิจัยด้านการสร้างคลังข้อมูลภาษาผู้เรียนเป็นการสร้างคลังข้อมูลภาษาโดยใช้งานเขียนของผู้เรียนภาษาต่างประเทศ เพื่อวิเคราะห์ข้อผิดพลาดด้านการใช้ภาษาต่างประเทศของผู้เรียน และเป็นที่สังเกตคือคลังข้อมูลของทั้ง 2 เรื่องนี้เป็นคลังข้อมูลภาษาต่างประเทศ ไม่พบงานวิจัยที่สร้างคลังข้อมูลภาษาผู้เรียนที่เป็นภาษาไทยของผู้เรียนชาวต่างชาติ

1.6 งานด้านการแปลภาษาด้วยเครื่อง

การวิจัยด้านการแปลโดยใช้คลังข้อมูลภาษาเป็นฐานส่วนใหญ่หมายถึงการวิจัยการแปลที่ใช้คลังข้อมูลสองภาษาคู่ขนาน เป็นหนึ่งในวิธีการวิจัยเชิงประจักษ์ที่สำคัญในด้านการศึกษากการแปลในปัจจุบัน จากการศึกษาพบจำนวนงานวิจัยด้านนี้ทั้งหมด 2 เรื่อง และทั้ง 2 เรื่องล้วนสังกัดสาขาวิชาการแปลภาษาอังกฤษและไทยของมหาวิทยาลัยธรรมศาสตร์ เนื้อหาสำคัญของงานวิจัยมีรายละเอียดดังต่อไปนี้

ศรัญญา โพธิกุล (2559) ทำการวิจัยเรื่อง “การทดสอบประสิทธิภาพของฐานข้อมูลศัพท์เฉพาะด้านในโปรแกรมช่วยแปล Memsources” งานวิจัยนี้ได้รวบรวมคำศัพท์เฉพาะและสร้างฐานข้อมูลคำศัพท์เฉพาะด้านการเจาะระบบ เพื่อทดสอบว่านักทดสอบเจาะระบบกับนักแปลทั่วไปใช้ฐานข้อมูลศัพท์เฉพาะด้านการเจาะระบบเหมือนหรือแตกต่างกันอย่างไร การนำฐานข้อมูลศัพท์เฉพาะด้านการเจาะระบบไปใช้ในโปรแกรมช่วยแปลช่วยให้แปลคำศัพท์เฉพาะได้อย่างคงเส้นคงวาหรือไม่ และเพื่อสอบถามความคิดเห็นของนักทดสอบเจาะระบบกับนักแปลทั่วไปต่อฐานข้อมูลศัพท์เฉพาะด้านการเจาะระบบ การวิจัยพบว่า นักทดสอบเจาะระบบนิยมใช้คำแปลของตนมากกว่าใช้ฐานข้อมูลศัพท์ฯ ซึ่งส่งผลถึงเรื่องความคงเส้นคงวาของการแปลเพราะมีการใช้คำแปลที่หลากหลายทำให้ความคงเส้นคงวาในกลุ่มนักทดสอบเจาะระบบมีน้อยกว่ากลุ่มนักแปลทั่วไป ขณะที่นักแปลทั่วไปนิยมใช้ฐานข้อมูลศัพท์ฯ และงานแปลมีความคงเส้นคงวามากกว่า และผู้ประเมินทั้งสองกลุ่มเห็น



ว่าการใช้ฐานข้อมูลศัพท์ฯ ช่วยเพิ่มประสิทธิภาพการทำงานของนักแปล สะดวกในการใช้งาน แต่ควรขยายฐานข้อมูลศัพท์โดยการเพิ่มคำศัพท์เฉพาะและคำศัพท์ไม่เฉพาะให้มากขึ้น รวมถึงเพิ่มความหมายของคำศัพท์แต่ละคำเข้าไปด้วย

กัญญภัค สาสงเคราะห์ (2562) ทำการศึกษาเรื่อง “การศึกษาเชิงเปรียบเทียบการแปลเว็บไซต์โรงแรมระหว่างมนุษย์และเครื่องมือแปลภาษา Google Translate จากภาษาไทยเป็นภาษาอังกฤษ” เพื่อมุ่งศึกษาความแตกต่างของการแปลเว็บไซต์โรงแรมระหว่างมนุษย์และโปรแกรมแปลภาษา Google Translate (GT) ในการแปลข้อความประเภทบอกกล่าวและข้อความจูงใจ และวิเคราะห์ความแตกต่างของการแปลในระดับคำ ระดับประโยค และระดับข้อความ รวมถึงเปรียบเทียบความเป็นธรรมชาติในภาษาแปล โดยเลือกข้อมูลต้นฉบับภาษาไทย (SL) จากเว็บไซต์โรงแรมไทยทั้งหมด 29 แห่ง และศึกษาภาษาปลายทางจากคลังข้อมูลภาษาอังกฤษ 2 แหล่ง ได้แก่ คลังข้อมูลจากเว็บไซต์ของโรงแรม (HT) และคลังข้อมูลจาก GT ผลจากการศึกษาพบว่า HT และ GT ใช้กลวิธีแปลแบบตรงตัวมากที่สุด แต่ HT ปรากฏกลวิธีการแปลที่หลากหลายมากกว่าและมีสากลลักษณะของภาษาแปลมากกว่า GT โดยเฉพาะการแปลข้อความจูงใจ เนื่องจาก HT สามารถปรับบทแปลได้ทุกระดับ ในขณะที่ GT ยังไม่สามารถปรับบทแปลในระดับข้อความได้ และยังมีข้อผิดพลาดที่เกิดจากการแบ่งประโยคผิดปรากฏอยู่ การปรับบทแปลและการแบ่งประโยคที่ต่างกันนี้จึงเป็นปัจจัยหลักที่ทำให้ภาษาปลายทางของ GT และ HT ต่างกันอย่างชัดเจน โดยที่ (1) GT สามารถแปลคำทั่วไปได้ถูกต้องใกล้เคียงกับ HT แต่ยังมีข้อจำกัดด้านถ่ายทอดคำทางวัฒนธรรม และ HT มีการหลากคำที่หลากหลายกว่า GT รวมถึงมีการใช้คำสร้างจินตภาพ ระดับคำที่เป็นทางการ และเฉพาะทางมากกว่า GT (2) HT และ GT มีโครงสร้างอนุประโยคของประโยคความซ้อนใกล้เคียงกัน แต่ HT สามารถปรับโครงสร้างประโยคคำสั่งในภาษาปลายทางได้หลากหลายกว่า GT ซึ่ง GT ยังมีปัญหาด้านไวยากรณ์ที่เกิดจากการแบ่งประโยคผิดอย่างชัดเจน (3) GT ไม่สามารถปรับการถ่ายทอดอนุวัจน์และเครื่องมือเชื่อมโยงความด้วยการเพิ่มความ ละความ สลับข้อความ หรือเปลี่ยนประเภทเครื่องมือเชื่อมโยงความได้เช่นเดียวกับ HT (4) HT สามารถปรับคำบ่งชี้บุคคลและคำปรากฏร่วมให้เข้ากับภาษาปลายทางได้อย่างเป็นธรรมชาติมากกว่า GT (5) อย่างไรก็ตาม HT และ GT ยังใช้คำปรากฏร่วมที่ไม่เป็นธรรมชาติมากนักเมื่อเทียบกับคลังข้อมูลเฉพาะทางที่เขียนโดยเจ้าของภาษา (L1) และคลังข้อมูล HT และ GT มีความสอดคล้องกับคลังข้อมูล L1 ในระดับหมวดหมู่ความหมายเท่านั้น แต่หากพิจารณาความถี่อันดับต้น การใช้คำปรากฏร่วมในคลังข้อมูลภาษาแปลยังแตกต่างกับคลังข้อมูล L1 ค่อนข้างชัดเจน



จากงานวิจัยทั้ง 2 เรื่องที่กล่าวมา จะเห็นได้ว่างานวิจัยด้านนี้มุ่งเปรียบเทียบความแตกต่าง การระหว่งการแปลด้วยเครื่องและการแปลด้วยมนุษย์ และความแตกต่างการระหว่งการแปลด้วย คลังข้อมูลที่แตกต่างกัน ส่วนผลการวิจัยสามารถสรุปได้ว่าการใช้ฐานข้อมูลศัพท์ฯ สามารถช่วยเพิ่ม ประสิทธิภาพการทำงานของนักแปล สะดวกในการใช้งาน แต่มีการใช้คำปรากฏร่วมที่ไม่เป็น ธรรมชาติ ควรขยายฐานข้อมูลศัพท์โดยการเพิ่มคำศัพท์เฉพาะและคำศัพท์ไม่เฉพาะให้มากขึ้น รวมถึงเพิ่มความหมายของคำศัพท์แต่ละคำเข้าไปด้วย สำหรับความถูกต้องของการแปล คลังข้อมูล เฉพาะทางที่เขียนโดยเจ้าของภาษามีความถูกต้องมากที่สุด ตามด้วยคลังคลังข้อมูลเฉพาะทาง และ คลังข้อมูลทั่วไปที่ไม่ใช่เฉพาะทาง

2 ผลการศึกษาด้านสถานภาพงานวิจัย

การวิจัยแบบอิงคลังข้อมูลภาษาสามารถครอบคลุมทุก ๆ ด้านของภาษาศาสตร์ รวมระดับ เสียง ระดับคำ ระดับวลี ระดับประโยค ระดับข้อความและวาทกรรม รวมภาษาพูดและภาษาเขียน รวมภาษาต่างกลุ่มภาษา เช่นภาษาต่างประเทศ ภาษาต่างวัย ภาษาต่างสมัย ภาษาต่างชาติพันธุ์ ภาษา ต่างท้องถิ่น รวมการทำพจนานุกรม การทำประมวลศัพท์ การแปลด้วยเครื่อง การสอนภาษา การ เรียนภาษาภาษาต่างชาติ การเปรียบเทียบภาษา การวิจัยด้านปริจเฉทวิเคราะห์ การวิจัยด้าน ภาษาศาสตร์สังคมวัฒนธรรม การวิจัยด้านภาษาศาสตร์จิตวิทยา ฯลฯ ระหว่างปี 2542-2563 งานวิจัยแบบอิงคลังข้อมูลภาษาภายในประเทศไทยมีการพัฒนาและก้าวหน้าอย่างเห็นได้ชัด แต่ก็มี ข้อจำกัดหรืออุปสรรคบางประการ จากการศึกษาเนื้อหาสำคัญของงานวิจัยทั้งหมด 52 เรื่อง สามารถสรุปภาพรวมของสถานภาพงานวิจัยของประเทศไทยได้ดังต่อไปนี้

2.1 เนื้อหาของงานวิจัยขาดความสมดุล งานวิจัยแบบอิงคลังข้อมูลของไทยนิยมทำการวิจัย ด้านการศึกษาคำศัพท์ อรรถศาสตร์ ไวยากรณ์และงานวิจัยด้านการเรียนการสอนภาษา โดย งานวิจัย 2 ประเภทนี้มีจำนวนมากที่สุดอยู่ที่ร้อยละ 50 และร้อยละ 21.15 แต่งานวิจัยประเภทอื่น มีจำนวนที่น้อยกว่า 2 ประเภทนี้เป็นอย่างมาก ได้แก่งานวิจัยด้านการเปรียบเทียบภาษา (ร้อยละ 11.54) งานวิจัยด้านการประมวลศัพท์ (ร้อยละ 9.61) งานวิจัยด้านคลังข้อมูลภาษาผู้เรียน (ร้อยละ 3.85) และงานวิจัยด้านการแปลภาษาด้วยเครื่อง (ร้อยละ 3.85) ซึ่งสถานภาพเช่นนี้ถือว่าเนื้อหาของ งานวิจัยขาดความสมดุล

2.2 งานวิจัยร้อยละ 92.65 มุ่งศึกษาภาษาระดับคำ ได้แก่งานวิจัยด้านการศึกษารหัสศาสตร์ และไวยากรณ์ระดับคำ การสอนภาษาระดับคำ การสร้างรายการคำศัพท์ การเปรียบเทียบ



ภาษาระดับคำ การศึกษาการใช้คำที่ไม่ถูกต้องของผู้เรียนชาวต่างชาติ ไม่พบงานวิจัยที่ศึกษา อรรถศาสตร์และไวยากรณ์ระดับเสียง ระดับวลี ระดับประโยค ระดับข้อความหรือวาทกรรม

2.3 งานวิจัยที่พบล้วนเป็นการวิจัยภาษาเขียน ไม่มีการวิจัยภาษาพูด

2.4 งานวิจัยที่พบไม่มีงานวิจัยภาษาต่างกลุ่มภาษา ได้แก่ภาษาต่างประเทศ ภาษาต่างวัย ภาษา-ต่างสมัย ภาษาต่างชาติพันธุ์ และภาษาต่างท้องถิ่น

2.5 ภาษาของตัวคลังข้อมูลขาดความสมดุล ได้แก่ งานวิจัยที่จำเป็นต้องสร้างคลังข้อมูลด้วย ผู้วิจัยเอง ส่วนมากเป็นคลังข้อมูลภาษาต่างประเทศ เช่นคลังข้อมูลภาษาอังกฤษ คลังข้อมูลภาษาจีน คลังข้อมูลภาษาญี่ปุ่น พบงานวิจัยเพียง 3 เรื่องเท่านั้นที่สร้างคลังข้อมูลภาษาไทย

จากที่กล่าวมาจะเห็นได้ว่า เนื้อหาของงานวิจัย ภาษาของคลังข้อมูลที่ใช้ในงานวิจัยขาด ความสมดุลและความหลากหลาย

สรุปผลและอภิปรายผล

หลายปีที่ผ่านมา การสร้างคลังข้อมูลและงานวิจัยที่ประยุกต์ใช้คลังข้อมูลในประเทศไทยได้รับการพัฒนาอย่างเห็นได้ชัด และประสบความสำเร็จมากมาย งานวิจัยแบบอิงคลังข้อมูลภาษามีจำนวนเพิ่มขึ้นเรื่อย ๆ จากการศึกษาครั้งนี้พบว่างานวิจัยแบบอิงคลังข้อมูลภาษาของประเทศไทยมีอยู่ 6 ประเภท และวิเคราะห์จากเนื้อหาสำคัญของงานวิจัยทั้งหมด 52 เรื่อง พบว่าสถานภาพงานวิจัยขาดความสมดุลและความหลากหลาย สถานภาพเช่นนี้สามารถอภิปรายเหตุผลจากตัวคลังข้อมูลและเทคโนโลยีหรือโปรแกรมคลังข้อมูลภาษาได้ดังต่อไปนี้

ภาษาไทยเป็นภาษาที่ไม่มีช่องว่างระหว่างคำ มีคำศัพท์ขนาดใหญ่ มีโครงสร้างทางไวยากรณ์ที่ซับซ้อน และประโยคมีการเชื่อมต่อกันอย่างไม่กระชับ ซึ่งทำให้เกิดความยากลำบากในการแบ่งคำ โปรแกรมคลังข้อมูลภาษาไทยที่มีอยู่แล้วยังไม่ถูกต้องแม่นยำในเรื่องการแบ่งคำ ยังต้องอาศัยการปรับปรุงแก้ไขจากมนุษย์อีกครั้ง แต่ภาษาต่างประเทศ โดยเฉพาะกลุ่มภาษาละตินที่มีช่องว่างระหว่างคำ ทำให้เรื่องการแบ่งคำถูกต้องกว่าภาษาไทยเป็นอย่างมาก ทำให้การวิจัยด้านภาษาต่างประเทศ โดยเฉพาะภาษาอังกฤษสามารถทำวิจัยได้ไกลกว่าและหลากหลายมากกว่า ทั้งนี้จึงทำให้งานวิจัยที่จำเป็นต้องสร้างคลังข้อมูลด้วยผู้วิจัยเอง ส่วนมากเป็นคลังข้อมูลภาษาต่างประเทศ และทำให้งานวิจัยส่วนมากเป็นการศึกษาระดับคำที่วิเคราะห์ความถี่และคำปรากฏร่วม ซึ่งตรงกับแนวคิดของ Baker (2010, pp. 1-2) ที่ว่าภาษาศาสตร์คลังข้อมูล (Corpus linguistics) เป็นการศึกษาภาษาที่ใช้ข้อมูลภาษาจริงจากฐานข้อมูลที่บันทึกไว้ในรูปแบบอิเล็กทรอนิกส์ทำให้



เราสามารถสืบค้นและศึกษาภาษาในเชิงสถิติได้มากขึ้น เช่น ดูความถี่ของการปรากฏ (frequency) ของปรากฏการณ์ทางภาษาหนึ่ง ๆ หรือคุณลักษณะและการแปรของภาษาตามตัวแปรด้านบริบทการใช้ เป็นต้น โดยรูปแบบการวิเคราะห์ที่เป็นที่นิยมรูปแบบหนึ่งคือการวิเคราะห์คำปรากฏร่วม และความถี่ของการปรากฏ โดยมีแนวคิดที่ว่าคำที่ปรากฏร่วมกันในความถี่สูงจะส่งผลต่อความหมายของคำนั้น ๆ (อ้างในอิงยศ กันจินะ (2559)) การพัฒนาต่อไปของงานวิจัยแบบอิงคลังข้อมูลภาษาไทยยังต้องการเทคโนโลยีหรือโปรแกรมคลังข้อมูลที่มีประสิทธิภาพและสะดวกยิ่งขึ้น เพื่อให้ นักวิจัยต่าง ๆ สามารถประยุกต์ใช้ในการทำวิจัยอย่างสะดวก มีประสิทธิภาพ ลึกซึ้งกว่าและละเอียดกว่า

นอกจากที่กล่าวมาแล้ว คลังข้อมูลภาษาไทยขนาดใหญ่ที่เผยแพร่ตัวคลังข้อมูลแก่สาธารณชน มีเพียงคลังข้อมูลภาษาไทยแห่งชาติ (Thai National Corpus) แหล่งเดียว แม้ว่าที่ผ่านมาจะมีหน่วยงานและสถาบันการศึกษาหลายแห่ง เห็นถึงประโยชน์และได้จัดสร้างคลังข้อมูลภาษาไทยขึ้นเพื่อใช้ในงานต่าง ๆ ของตน แต่เนื่องจากปัญหาเรื่องลิขสิทธิ์ทำให้ไม่สามารถเผยแพร่ข้อมูลที่ทำได้ มีเพียงคลังข้อมูลอรรถคดีของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) เท่านั้นที่ได้เผยแพร่แก่สาธารณะ เพราะเป็นข้อมูลที่ไม่มีลิขสิทธิ์ แต่ก็ยังเป็นข้อมูลขนาดเล็ก (4 แสนคำ) และเป็นงานเขียนวิชาการประเภทเดียว นอกจากนี้ ก็มีคลังข้อมูลภาษาไทยที่ภาควิชาภาษาศาสตร์ จุฬาฯ ได้รวบรวมจากแหล่งต่าง ๆ เช่น หนังสือพิมพ์ สุนทรพจน์ รายงานข่าว บทความวิชาการ นิยายและเรื่องสั้น และเปิดให้บริการสืบค้นข้อมูลภาษาไทยผ่านอินเทอร์เน็ต (<http://ling.arts.chula.ac.th/ThaiConc/>) แต่ก็ยังเป็นข้อมูลที่จัดเก็บตามสะดวก ไม่ครอบคลุมการใช้ภาษาไทยในลักษณะต่าง ๆ อย่างเป็นระบบ และลิขสิทธิ์ของงานเขียนต่าง ๆ ยังคงเป็นของสำนักพิมพ์และผู้เขียน จึงไม่สามารถเผยแพร่ตัวคลังข้อมูลทั้งหมดแก่สาธารณชนได้ (อ้างในโครงการ Thai National Corpus <https://www.arts.chula.ac.th/ling/tnc/>)

อย่างไรก็ตาม แม้ว่าคลังข้อมูลภาษาไทยคลังข้อมูลภาษาไทยแห่งชาติมีมากกว่าสามสิบล้านคำ แต่เทียบกับคลังข้อมูลภาษาต่างประเทศที่มีขนาดอยู่ที่หลายร้อยล้านคำ (เช่น คลังข้อมูลของประเทศที่พูดภาษาอังกฤษ, คลังข้อมูลภาษาอังกฤษร่วมสมัยแบบอเมริกัน) การสร้างคลังข้อมูลของภาษาไทยยังจำเป็นต้องขยายขนาดมากขึ้น ในขณะเดียวกัน เนื้อหาของตัวคลังข้อมูลภาษาไทยแห่งชาติขาดความสมบูรณ์ หมวดหมู่ของประเภทงานเขียนมีเพียง 6 หมวด (ได้แก่ งานวิชาการ กึ่งวิชาการ เรื่องแต่ง หนังสือพิมพ์ กฎหมาย และอื่น ๆ) ไม่มีงานเขียนประเภทอื่น ๆ ไม่มีคลังข้อมูลภาษาพูด และภาษาของกลุ่มภาษาหลายกลุ่มยังไม่ได้รวบรวมในคลังข้อมูลอย่างเป็นระบบ เช่น ภาษาต่างประเทศ ภาษาต่างวัย ภาษาต่างสมัย ภาษาต่างชาติพันธุ์ ภาษาต่างชาติ ภาษาต่างท้องถิ่น ฯลฯ



ซึ่งการจัดสร้างคลังข้อมูลเหล่านี้ มีความสำคัญอย่างยิ่งสำหรับการพัฒนาของคลังข้อมูลภาษา ปัจจัยเหล่านี้จึงทำให้สถานภาพงานวิจัยที่ประยุกต์ใช้คลังข้อมูลขาดความสมดุลและความหลากหลาย

ตลอดหลายปีที่ผ่านมา การพัฒนาภาษาศาสตร์คลังข้อมูลภาษาต่างประเทศมีแนวโน้มจากการวิจัยปรากฏการณ์ของภาษาเองกลายเป็นการวิจัยแบบสหวิทยาการกับสาขาอื่นที่เกี่ยวข้อง โดยเฉพาะอย่างยิ่งกับภาษาศาสตร์สังคม (Beaugrande, 2008; Anderson, 2008), ภาษาศาสตร์แห่งความรู้ความเข้าใจ (Vukovic, 2007), ภาษาคอมพิวเตอร์ (Orasan et al., 2007) และสาขาวิชาอื่นที่เป็นการวิจัยสหวิทยา ในขณะเดียวกัน นักวิชาการต่างชาติจำนวนมากได้เริ่มสนใจทำการวิจัยเชิงปฏิบัติของคลังข้อมูล ได้แก่ การประยุกต์ใช้คลังข้อมูลกับการดึงข้อมูลตามสาขาวิชาเฉพาะ (Braschler, 2004; Miangah & Nezarat, 2010) การประเมินผลงานภาษา (Deane & Quinlan, 2010; Shams, 2010; Islam & Rajon, 2010) และอื่น ๆ ในอนาคต การวิจัยภาษาศาสตร์คลังข้อมูลภาษาของไทยก็คงเดินทางไปยังทิศทางเหล่านี้เช่นกัน



เอกสารอ้างอิง

- กะรัตเพชร คงรอด. (2560). การวิเคราะห์ข้อผิดพลาดด้านการใช้คำในการเขียนภาษาอังกฤษเชิงวิชาการของนักศึกษาไทยระดับปริญญาตรี. สาขาวิชาการสอนภาษาอังกฤษ ภาควิชาหลักสูตรและวิธีสอน มหาวิทยาลัยศิลปากร.
- กัญญาภักดิ์ สาสงเคราะห์. (2562). การศึกษาเชิงเปรียบเทียบการแปลเว็บไซต์โรงแรมระหว่างมนุษย์และเครื่องมือแปลภาษา Google Translate จากภาษาไทยเป็นภาษาอังกฤษ. วิทยานิพนธ์ สาขาวิชาการแปลภาษาอังกฤษและไทย คณะศิลปศาสตร มหาวิทยาลัยธรรมศาสตร์.
- ขวัญชนิตร์ สุทธิสารสุนทร. (2546). กระบวนการทำประมวลศัพท์หมวดพยานิชย์อิเล็กทรอนิกส์. วิทยานิพนธ์ สาขาวิชาการแปลและล่าม คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
- จิตสุตา บุญรัตน์ , คเชนทร์ ตัญศิริ, มุกซ์ดา สุขธาราจารย์. (2563). การศึกษาคุณสมบัติทางความหมายของการปรากฏร่วมของคำว่า “ติด” ในภาษาไทยโดยใช้คลังข้อมูลภาษา. วารสารวณะ 8(1).
- จุฬารัตน์ คำน้อย, หลิวซู, จิ้นหลิง. (2561). การสร้างคลังข้อมูลและวิเคราะห์การเขียนตัวอักษรจีนผิดของนักศึกษาไทย ระดับอุดมศึกษา กรณีศึกษาสำนักวิชาจีนวิทยา มหาวิทยาลัยแม่ฟ้าหลวง. สำนักจีนวิทยา มหาวิทยาลัยแม่ฟ้าหลวง เชียงราย.
- ชาตบุษย์ นิสระรัตน์. (2551). คำปรากฏร่วมจำเพาะและอรรถสัมผัสของคำกริยา “เกิด” ในภาษาไทย-การศึกษาเชิงภาษาศาสตร์คลังข้อมูล. วารสารภาษาและภาษาศาสตร์.
- ชุติมา เตชะพิรุฬห์. (2550). ประมวลศัพท์เรื่องเครื่องหมายการค้า. สารนิพนธ์ สาขาวิชาการแปลและการล่าม คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- ชื้อณี เหยา. (2561). ระบบโมทัศน์ “ใจ” ในภาษาไทยเปรียบเทียบกับภาษาจีน: การศึกษาตามแนวภาษาศาสตร์ปริชาน. ดุษฎีนิพนธ์ สาขาภาษาศาสตร์ มหาวิทยาลัยธรรมศาสตร์.
- ธารทอง แจ่มไพบูลย์. (2559). พัฒนาการของคำศัพท์ที่มีความถี่ในการปรากฏสูงในหนังสือเรียนวิชาภาษาไทย: การศึกษาตามแนวทางภาษาศาสตร์คลังข้อมูล. ดุษฎีนิพนธ์ สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- นพวรรณ เมืองแก้ว. (2556). คำแสดงการขอร้องในภาษาไทย : การศึกษาตามแนวทฤษฎีการกลายเป็นคำไวยากรณ์. กรุงเทพมหานคร: จุฬาลงกรณ์มหาวิทยาลัย.



- ปราริสสา เหล่าวนิชย์. (2552). *ประมวลศัพท์เรื่องการสำรวจและผลิต การขนส่งและรูปแบบการใช้เชื้อเพลิงก๊าซธรรมชาติ*. วิทยานิพนธ์ระดับปริญญาโท สาขาวิชาการแปลและการล่าม คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- ปิยะพงษ์ เลาศรีรัตนชัย. (2562). *การสร้างรายการคำศัพท์เพื่อธุรกิจบริการและบทเรียนคำศัพท์ออนไลน์*. สาขาวิชาภาษาศาสตร์ คณะมนุษยศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ.
- ปิยะวดี คำสุวรรณ, คเชนทร์ ตัญศิริ, อุมารณ สักขมาน. (2562). การศึกษาคำว่า “ตาม” ในภาษาไทย. *วรรณวิทัศน์*, 19(2), 179-207.
- ผไทมาศ ประมวล , พุทธชาติ โปธิบาล , คเชนทร์ ตัญศิริ. (2563). *เครือข่ายความหมาย ‘นอก’ ในภาษาไทย*. สาขาภาษาศาสตร์ประยุกต์ คณะมนุษยศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.
- พิรุฬห์ ปิยมพวงศ์. (2559). *คุณสมบัติทางวากยสัมพันธ์และอรรถศาสตร์ของตัวบ่งชี้ผู้รับประโยชน์ “เพื่อ” และ “ให้” ในภาษาไทย*. วิทยานิพนธ์ สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- ยิ่งยศ กันจินะ. (2559). *การศึกษาคำหมายของคำว่า “พม่า” ด้วยคำปรากฏร่วมจากคลังข้อมูลภาษาไทยแห่งชาติ*. สาขาวิชาภาษาศาสตร์ สถาบันวิจัยภาษาและวัฒนธรรมเอเชีย มหาวิทยาลัยมหิดล.
- รังสิมา รุ่งเรือง. (2560). *ความหมายเปรียบเทียบของสัตว์ในภาษาไทยและภาษาอินโดนีเซีย*. วิทยานิพนธ์ สาขาวิชาภาษาศาสตร์เพื่อการสื่อสาร ภาควิชาภาษาอังกฤษและภาษาศาสตร์ คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์.
- วิโรจน์ อรุณมานะกุล. (2553). *ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้*. กรุงเทพฯ: โครงการเผยแพร่ผลงานวิชาการ คณะอักษรศาสตร์จุฬาลงกรณ์มหาวิทยาลัย.
- ศรัณญา โพธิกุล. (2559). *การทดสอบประสิทธิภาพของฐานข้อมูลศัพท์เฉพาะด้านในโปรแกรมช่วยแปล Memsource*. วิทยานิพนธ์สาขาวิชาการแปลภาษาอังกฤษและไทย คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์.
- อภิชาติ คานนิม. (2561). *การศึกษามโนทัศน์การเคลื่อนที่เข้าในภาษาไทยและภาษาญี่ปุ่น*. ดุษฎีนิพนธ์ สาขาวิชาภาษาศาสตร์ คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์, คลังข้อมูลภาษาไทยแห่งชาติ <https://www.arts.chula.ac.th/ling/tnc/structure/>



- Anderson, W. (2008). *Corpus Linguistics in the UK: Resources for Sociolinguistic Research. Language and Linguistics Compass*, 2008 (2): 113-125.
- Beaugrande, R. (2008). Linguistics, Sociolinguistics and Corpus Linguistics: Ideal Language versus Real Language. *Journal of Sociolinguistics*, 2008 (1): 128-139.
- Braschler, M. & Schauble P. (2004). Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. *Information Retrieval*, (3): 27-39.
- Deane, P. & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2010 (2): 151-177.
- Islam, M. R. & Rajon, S. A. (2010). Design and analysis of an effective corpus for evaluation of Bengali text compression schemes. *Journal of Computers*, 2010 (1): 59-68.
- Orason, C. et al. (2007). Corpora for computational linguistics. *Ilha Do Desterro: A Journal of English Language. Literature in English and Cultural Studies*, 2007, 52(1): 65-102.
- Vukovic, N.T. (2007). Corpora in cognitive linguistics. *Contemporary Linguistics*, 2007, (1): 102-107.
- Zhao Jiaoyue. (2558). การศึกษาความแตกต่างการใช้คำศัพท์ และวิธีการสอนภาษาจีน เกี่ยวกับ
ปาลง หนอน เต่า ในคลังคำศัพท์. วิทยานิพนธ์ศิลปศาสตรมหาบัณฑิต (การสอนภาษาจีน)
มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ.