งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาบริบทที่สามารถเป็นบริบทบ่งบอกการปรากฏของชื่อเฉพาะภาษาไทย สำหรับนำไปใช้ในการพัฒนางานด้านภาษาศาสตร์คอมพิวเตอร์ โดยศึกษาจากคลังข้อมูลภาษาที่มีการกำกับชื่อเฉพาะ ผู้วิจัยได้ทำการจัดสร้างคลังข้อมูลภาษาขึ้น โดยรวบรวมข้อความภาษาไทยจากหนังสือพิมพ์กรุงเทพธุรกิจและนิตยสารสกุลไทยแล้วทำการกำกับชื่อเฉพาะและบริบทบ่งบอกตามแนวทางการกำกับชื่อเฉพาะของ TEI

ผลการศึกษาชื่อเฉพาะทั้งหมด 2,547 ชื่อ พบว่าชื่อเฉพาะที่ปรากฏบริบทบ่งบอกคิดเป็น 63% ชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอกคิดเป็น 37% บริบทที่สามารถบ่งบอกชื่อเฉพาะภาษาไทยได้แก่ (1) คำที่ปรากฏติดกับชื่อเฉพาะหรือคำบ่งบอก (2) การเว้นวรรค และ (3) ตัวบ่งบอกระดับปริจเฉท คำบ่งบอกและการเว้นวรรคเป็นตัวบ่งบอกที่ปรากฏกับชื่อเฉพาะโดยส่วนใหญ่ คำบ่งบอกมักจะปรากฏติดอยู่หน้าชื่อเฉพาะ ในการศึกษานี้พบคำบ่งบอก 161 คำ ซึ่งสามารถจำแนกเป็นกลุ่มตามอรรถลักษณ์ร่วมในกลุ่มนั้นได้ทั้งสิ้น 19 กลุ่ม คำบ่งบอกยังสามารถบ่งชี้ประเภทของชื่อเฉพาะโดยชื่อเฉพาะประเภทหนึ่งจะปรากฏร่วมกับคำบ่งบอกกลุ่มเฉพาะกลุ่มหนึ่ง ประเภทของชื่อเฉพาะที่พบในการศึกษานี้มี 6 ประเภท ได้แก่ ชื่อคน ชื่อองค์กร ชื่อสถานที่ ชื่อกลุ่มคนร่วมเชื้อสาย ชื่อกฎหมาย และชื่ออื่นๆ นอกจากนี้ยังพบว่าคำบ่งบอกมีขีดความสามารถในการบ่งบอกชื่อเฉพาะแตกต่างกันแม้ว่าจะเป็นคำบ่งบอกในกลุ่มเดียวกัน ส่วนการเว้นวรรคปรากฏในตำแหน่งส่วนหน้าหรือส่วนหลังของชื่อเฉพาะและเป็นบริบทบ่งบอกที่มีความสัมพันธ์กับคำบ่งบอก การวิจัยพบว่าชื่อเฉพาะจำนวน 71% มีการเว้นวรรคร่วมด้วย และ 80% ของชื่อเฉพาะที่มีคำบ่งบอกจะมีการเว้นวรรคด้วย หากพิจารณาการเว้นวรรคร่วมกับคำบ่งบอกจะทำให้คำบ่งบอกมีขีดความสามารถในการบ่งบอกชื่อเฉพาะสูงขึ้น และตัวบ่งบอกระดับปริจเฉท แม้จะพบเป็นจำนวนน้อยแต่มีรูปแบบที่ชัดเจน สามารถบ่งบอกการปรากฏชื่อเฉพาะที่ไม่ปรากฏคำบ่งบอกได้ โดยส่วนใหญ่พบว่าเป็นลักษณะอ้างตามชื่อที่ปรากฏก่อนในกรณีชื่อเฉพาะ 931 ชื่อที่ไม่ปรากฏบริบทบ่งบอกพบว่า เป็นชื่อเฉพาะที่มีโครงสร้างของรูปภาษาในลักษณะทับศัพท์ภาษาต่างประเทศ หรือเป็นชื่อเฉพาะที่ผู้ใช้ภาษามีภูมิความรู้ร่วมกัน จึงเอื้อต่อการนำชื่อเฉพาะนั้นไปใช้โดยไม่ต้องมีตัวบ่งบอกในการระบุว่ารูปภาษานั้นเป็นชื่อเฉพาะ ในการศึกษาครั้งนี้ยังพบว่า นอกจากชื่อเฉพาะจะปรากฏเป็นหน่วยชื่อเฉพาะที่มีคุณสมบัติเฉพาะประจำรูปแล้ว ยังสามารถปรากฏเป็นส่วนประกอบหนึ่งของรูปภาษาอื่นที่ไม่ใช่ชื่อเฉพาะอีกด้วย

This study is a computational linguistic study of context clues. It aims to analyse context clues that could be used to identify proper names in Thai. A corpus is collected from Bangkok Business newspaper and Sakulthai magazine. Proper names and context clues are manually tagged by using the Text Encoding Initiative (TEI)'s guideline.

Based on the studying of 2,547 proper names, 63 percents of proper names occur with context clues, while 37 percents do not have any context clues. The context clues consist of (1) clue words, (2) spaces and (3) discourse context. Clue words are normally in front of proper names. 161 clue words are found in this study. They can be grouped in according to their semantic features into 19 groups. Each group of clue words can be used to indicate types of proper names. Six types of proper names are identified in this study, namely person, organization, place, affinity, law, and miscellaneous. Furthermore, different clue words tend to have different predictive power in identifying proper names, even when they are in the same group. A space usually occurs together with a clue word. It can appear either immediately before or after the proper name. It is found that 71 percents of proper names occur with a space, and 80 percents of proper names that have clue words do occur with a space also. Most clue words have more predictive power when spacing is taking into consideration. When proper names do not occur with clue words or spaces, discourse context can be used to identify proper names. Discourse context here refers to anaphoric relation or cataphoric relation, but it is mostly anaphoric one. These proper names can be used without any clue words or spaces because it refers to the same discourse entity mentioned before. However, there are 931 proper names that cannot be identified by any context clues. These proper names are either foreign words or well-known names. Therefore, these proper names can be used without context clues. In addition, we also found that proper names can be used as a part of linguistic units that are not proper names.