ภาษาศาสตร์คอมพิวเตอร์

วิโรจน์ อรุณมานะกุล
ภาควิชาภาษาศาสตร์, คณะอักษรศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
24 สค. 2543 (Draft)

เป็นที่รู้กันดีว่า การที่มนุษย์เราจะสั่งให้คอมพิวเตอร์ทำงานอะไรนั้น เราจำเป็นต้องมีโปรแกรมที่เขียนขึ้นด้วยภาษาที่คอมพิวเตอร์เข้าใจได้ เช่น ภาษาแอสเซมบลี ภาษาซี ภาษาเบสิก ความฝันที่จะทำให้คอมพิวเตอร์สามารถพูดคุยติดต่อกับมนุษย์ด้วยภาษาของมนุษย์เองนั้นเป็นความฝันที่มนุษย์มีมานานแล้ว ดังที่เราจะเห็นความคิดเหล่านี้ปรากฏอยู่ในหนังสือหรือภาพยนตร์นิยายวิทยาศาสตร์ต่างๆอยู่เสมอในรูปของหุ่นยนตร์ที่สามารถพูดคุยติดต่อสื่อสารกับมนุษย์ได้เหมือนมนุษย์คนหนึ่ง เช่น คอมพิวเตอร์ที่ชื่อว่า HAL ในหนังสือและภาพยนตร์เรื่อง 2001: A Space Odyssey โดยที่ HAL เป็นคอมพิวเตอร์ประจำยานอวกาศที่สามารถพูดคุยกับนักบินอวกาศ สามารถทำความเข้าใจกับประโยคที่ได้ยิน และมีความคิดเป็นของตัวเอง HAL เป็นตัวอย่างของการประยุกต์เรื่องปัญญาประดิษฐ์ (artificial intelligence) ที่นอกจากจะทำให้คอมพิวเตอร์สามารถคิดใช้เหตุผล มีฐานความรู้ต่างๆและรับรู้เรื่องของโลกภายนอกได้แล้ว HAL ยังมีสามารถทางด้านภาษา สามารถฟังและรับรู้คลื่นเสียงที่ได้ยินว่าพูดถึงประโยคอะไร (speech recognition) สามารถเข้าใจประโยคที่ได้ยินว่าประกอบด้วยคำอะไรบ้างมีความหมายอย่างไร (natural language understanding) และสามารถถ่ายทอดความคิดที่ต้องการสื่อออกมาเป็นภาษาของมนุษย์ (natural language generation) และสามารถถ่ายทอดประโยคที่ต้องการออกมาในรูปของคลื่นเสียงที่มนุษย์สามารถได้ยินและรับรู้ได้ (speech synthesis)

ประโยชน์ที่ได้จากการทำให้คอมพิวเตอร์สามารถติดต่อสื่อสารกับมนุษย์ด้วยภาษามนุษย์เองนั้นชัดเจนในตัวเอง เพราะจะส่งผลให้การใช้งานคอมพิวเตอร์เป็นไปอย่างสะดวกมากขึ้น คอมพิวเตอร์จะสามารถเข้ามาช่วยในงานด้านต่างๆที่เกี่ยวข้องกับภาษาได้มากขึ้น เช่น เป็นเครื่องแปลภาษามนุษย์จากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง (machine translation) สามารถนำคอมพิวเตอร์มาช่วยตรวจและวิเคราะห์เอกสารต่างๆที่มีว่าเกี่ยวข้องกับเรื่องใด ทำให้สามารถใช้คอมพิวเตอร์เพื่อช่วยในการค้นคืนข้อมูล (information retrieval) ตามความต้องการของผู้ใช้ได้ หรือสามารถให้คอมพิวเตอร์ช่วยสรุปสาระและประเด็นสำคัญๆที่ปรากฏในเอกสารนั้นๆ (information extraction) เป็นต้น ศาสตร์ที่เกี่ยวข้องกับการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์นี้คือศาสตร์ที่เรียกว่า ภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics)

ภาษาศาสตร์คอมพิวเตอร์เกี่ยวข้องกับศาสตร์ใดบ้าง

ศาสตร์ที่ศึกษาทางด้านนี้เป็นแขนงวิชาหนึ่งที่เกี่ยวข้องกับศาสตร์หลายๆศาสตร์ เราเรียกศาสตร์ด้านนี้อีกชื่อหนึ่งว่าการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ศาสตร์ที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ ได้แก่ คอมพิวเตอร์ ภาษาศาสตร์ จิตวิทยา วิศวกรรมไฟฟ้า และสถิติ โดยที่ในทางคอมพิวเตอร์จะเน้นที่การศึกษาในเรื่องของระบบการประมวลผลภาษาธรรมชาติ (NLP) เรื่องของการแทนรูปความรู้ (knowledge

representation) เรื่องของเทคนิคต่างๆของการแจงส่วนประโยค เป็นต้น ในทางภาษาศาสตร์จะเน้นที่เรื่องของการวิเคราะห์องค์ประกอบต่างๆของภาษา ในทางวิศวกรรมไฟฟ้าจะสนใจในแบบจำลองต่างๆ ที่ใช้สำหรับระบบต่างๆ ทางด้านเสียง ไม่ว่าจะเป็นทางด้านการสังเคราะห์เสียง (speech synthesis) หรือการรู้จำเสียง (speech recognition) ในทางสถิติสนใจเรื่องของการประยุกต์ใช้ความรู้ทางสถิติในแบบจำลองภาษา (language model) ที่ใช้

ความคิดในเรื่องการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์นั้นมีมาตั้งแต่ในยุคแรก ๆ ของการใช้คอมพิวเตอร์ ในปี 1950 Alan Turing ซึ่งเป็นผู้บุกเบิกศาสตร์ด้านคอมพิวเตอร์ในระยะแรก ได้เสนอวิธีการที่จะใช้ทดสอบว่าคอมพิวเตอร์มีความสามารถทางภาษาหรือไม่ โดยให้มีการทดลองติดต่อสื่อสารผ่านทางหน้าจอเทอร์มินัลในห้องที่แยกจากกัน โดยที่ข้อความของคู่สนทนาที่เห็นปรากฏอยู่บนจอนั้นอาจจะมาจากการพิมพ์ของคนที่ใช้คอมพิวเตอร์ในอีกห้องหนึ่ง หรืออาจจะมาจากการสร้างขึ้นมาเองของคอมพิวเตอร์ก็ได้ ถ้าหากว่าผู้ทดลองไม่สามารถแยกแยะได้ว่า ผู้ที่เขาสนทนาอยู่ด้วยนั้นเป็นคอมพิวเตอร์ ก็ให้ถือว่าคอมพิวเตอร์เครื่องนั้นผ่านการทดสอบครั้งนี้ แนวการทดสอบแบบนี้เรียกว่าการทดสอบแบบทัวริง (Turing Test) แนวคิดของการทดสอบแบบนี้ นำไปสู่ข้อถกเถียงมากมายว่าการสังเกตจากรูปภาษาที่ปรากฏเพียงอย่างเดียวสามารถใช้เป็นข้อสรุปว่าคอมพิวเตอร์มีความสามารถทางภาษาเหมือนมนุษย์ได้จริงหรือไม่ มีกลุ่มคนที่แย้งว่าการที่คอมพิวเตอร์สามารถจัดการกับสัญลักษณ์ต่างๆที่ให้ไม่ได้หมายความว่าคอมพิวเตอร์มีความเข้าใจทางภาษาเกิดขึ้นจริง ซึ่งในที่นี้จะไม่กล่าวถึงข้อถกเถียงเหล่านี้ ผู้สนใจสามารถอ่านแนวคิดนี้ได้จากงานของ John R. Searle ที่อ้างถึงกรณีปัญหา "Chinese Room" ในบทความ "Minds, Brains, and Programs" (ใน The Behavioral and Brain Sciences, 1980, 3, 422-424)

ตัวอย่างโปรแกรมคอมพิวเตอร์ในยุคแรกที่ถูกออกแบบมาให้ดูเสมือนมีความสามารถทางภาษา สามารถโต้ตอบกับผู้ใช้ด้วยภาษาอังกฤษได้เหมือนกับเป็นการพูดคุยอยู่กับจิตแพทย์ คือโปรแกรมที่รู้จักกันในชื่อของ Eliza โปรแกรม Eliza นี้สามารถโต้ตอบกับมนุษย์ได้โดยไม่มีกระบวนการตีความหรือวิเคราะห์หาความหมายของประโยคเลย โปรแกรมโต้ตอบโดยการเปรียบเทียบรูปแบบของประโยคที่พบ (pattern matching) เช่น ถ้าพบกับประโยคที่อยู่ในรูปแบบ “I like X” โดยที่ X หมายถึงข้อความอะไรก็ได้ โปรแกรม Eliza ก็ถูกกำหนดให้ตอบกลับด้วยรูปแบบประโยคว่า “Can you tell me why do you like X?” เป็นต้น ตัวอย่างเช่น ถ้าผู้ใช้พิมพ์ประโยคว่า “I like chocolate” โปรแกรมจะตอบกลับมาว่า “Can you tell me why do you like chocolate?” โปรแกรมแบบ Eliza นี้ทำให้คอมพิวเตอร์ดูเสมือนว่าสามารถติดต่อสื่อสารกับมนุษย์ได้ แต่ก็เป็นไปในขอบเขตที่จำกัดตามรูปแบบที่กำหนดให้เท่านั้น สาเหตุที่ผู้ใช้บางคนรู้สึกว่าสามารถพูดคุยกับโปรแกรม Eliza ได้รู้เรื่อง น่าจะเป็นเพราะความสามารถของมนุษย์เราเองที่พยายามหาความหมายและเชื่อมโยงความหมายจากข้อความต่าง ๆ ที่ได้เห็นจากหน้าจอมากกว่าที่จะเกิดจากการที่คอมพิวเตอร์มีความสามารถวิเคราะห์และเข้าใจความหมายทางภาษาอย่างแท้จริง ดังนั้น ในศาสตร์ทางด้านภาษาศาสตร์คอมพิวเตอร์นี้ เราจึงต้องการระบบการประมวลผลที่สลับซับซ้อนมากกว่าการเปรียบเทียบรูปแบบภาษาอย่างง่ายๆแบบโปรแกรม Eliza เราต้องการทำให้คอมพิวเตอร์สามารถประมวลผลภาษาได้อย่างแท้จริง สามารถแยกแยะได้ว่าประโยคที่ได้ยินหรือป้อนเข้าไปนั้นประกอบด้วยคำอะไรบ้าง มีความสัมพันธ์ทางวากยสัมพันธ์อย่างไร และเข้าใจว่าความหมายที่ผู้พูดตั้งใจจะสื่อคืออะไร

การที่เราจะทำให้คอมพิวเตอร์มีความสามารถทางภาษาเหมือนมนุษย์นั้น ก็มีกระแสความคิดหลักอยู่สองแบบ ในแบบแรกนั้นเชื่อว่า เราจำเป็นต้องเข้าใจกลไกความสามารถทางภาษาที่มนุษย์มีอยู่ว่าเป็นอย่างไรก่อน เราจึงจะสามารถสร้างแบบจำลองอันนั้นให้กับคอมพิวเตอร์ได้ ส่วนวิธีคิดแบบที่สอง มองว่าเราไม่จำเป็นต้องจำลองการทำงานด้านภาษาของคอมพิวเตอร์ให้เหมือนกับกระบวนการที่เกิดขึ้นในสมองมนุษย์จริงๆ กระบวนการที่ใช้อาจเป็นกระบวนการเฉพาะที่เหมาะสำหรับใช้กับคอมพิวเตอร์และสามารถทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์และโต้ตอบสื่อสารกับมนุษย์ได้ก็เพียงพอแล้ว ในกลุ่มหลังนี้ มักจะยกอ้างกรณีเรื่องเทคโนโลยีการบิน โดยกล่าวว่ามนุษย์สามารถสร้างเครื่องบินให้บินได้โดยไม่ต้องอาศัยกลไกการบินแบบที่ต้องกระพือปีกเหมือนนก แต่ใช้หลักการของความดันอากาศของเครื่องร่อน ในขณะที่การประดิษฐ์เครื่องบินในยุคแรก ๆ โดยความพยายามที่จะเลียนแบบการบินของนกกลับไม่ประสบความสำเร็จ จึงไม่มีความจำเป็นใดใดที่จะต้องมุ่งทำให้คอมพิวเตอร์ประมวลผลภาษาในลักษณะเดียวกับที่มนุษย์กระทำ

ในที่นี้ เราจะยึดแนวคิดแบบแรกเป็นหลักก่อน คือ พยายามสร้างระบบการประมวลผลภาษาธรรมชาติที่จำลองแบบการประมวลผลภาษาที่เกิดขึ้นจริงในมนุษย์ เพราะอย่างน้อยก็จะได้ช่วยให้เราเข้าใจกลไกทางภาษาของมนุษย์เองได้ดีขึ้น ซึ่งการที่จะทำเช่นนี้ได้ เราจำเป็นต้องพยายามทำความเข้าใจเสียก่อนว่าภาษาคืออะไร ธรรมชาติของภาษามนุษย์เป็นอย่างไร ทำไมเราจึงใช้ภาษาติดต่อสื่อสารซึ่งกันและกันได้ เพื่อที่เราจะได้เข้าใจธรรมชาติของภาษามนุษย์มากขึ้น ก่อนที่จะสร้างแบบจำลองใดใดให้กับคอมพิวเตอร์

<u>ภาษาคืออะไร</u>

ภาษาคืออะไร หลายคนอาจตอบว่า คือสื่อที่มนุษย์ใช้เพื่อติดต่อสื่อสารกัน แต่หากถามต่อว่า ทำไมคนเราถึงสามารถใช้ภาษาได้ สามารถพูด ฟัง เขียน หรือ อ่านได้ กระบวนการและขั้นตอนที่ใช้ในการเข้าใจภาษามีรายละเอียดอย่างไรบ้าง คนส่วนมากอาจตอบไม่ได้ เรารู้แต่เพียงว่าเรามีความสามารถทางภาษา สามารถอ่าน ฟัง หรือพูด เพื่อติดต่อสื่อสารกับคนอื่นๆได้ ภาษาเป็นความสามารถอย่างหนึ่งที่มีอยู่กับตัวเรา เราได้ใช้ภาษาอย่างเป็นปกติวิสัยจนเราแทบจะไม่นึกถึงว่ามีความสลับซับซ้อนมากเพียงใด ไม่มีสังคมมนุษย์ที่ไหนในโลกที่อยู่โดยไม่มีภาษา ใน Towel of Babel กล่าวไว้ว่า เดิมทีมนุษย์ในโลกนี้พูดภาษาเดียวกัน ต่อมามนุษย์ก็พยายามสร้างหอคอยสูงขึ้นไปเพื่อที่จะเอื้อมให้ถึงพระเจ้า พระเจ้าจึงลงโทษมนุษย์ด้วยการทำลายหอคอยอันนั้น และสาปให้มนุษย์แต่ละกลุ่มพูดคนละภาษา เพื่อไม่ให้มนุษย์สื่อสารกันได้รู้เรื่อง ไม่สามารถแลกเปลี่ยนความรู้ซึ่งกันและกัน อันจะนำไปสู่การสร้างสมรวบรวมพลังอำนาจ (collective power) ที่จะท้าทายต่ออำนาจของพระเจ้าได้อีก

คนจำนวนมากเชื่อกันว่าภาษาเป็นเรื่องที่เรียนรู้จากสังคมเหมือนอย่างการเรียนรู้วัฒนธรรมต่างๆในสังคม เช่น รู้ว่าจะต้องยืนตรงเวลาได้ยินเพลงชาติไทย รู้ว่าจะต้องถอดรองเท้าก่อนเดินเข้าโบสถ์ แต่จริง ๆ แล้วไม่ใช่ ภาษาไม่ใช่สิ่งประดิษฐ์ทางวัฒนธรรม (language is not a culture artifact) หรือเป็นสิ่งที่มนุษย์ประดิษฐ์ขึ้น ภาษาเป็นคุณสมบัติดั้งเดิมที่ติดตัวมนุษย์มาแต่กำเนิด ที่เรียกว่าเป็น instinct หรือเป็น innate ผู้ที่เสนอแนวความคิดนี้ คือ Chomsky ซึ่งเป็นผู้ที่ทำให้เกิดการเปลี่ยนแปลงแนวการศึกษาภาษาศาสตร์อย่างมาก (จากที่เคยมองภาษาและศึกษาภาษาตามแนวทางความเชื่อแบบพฤติกรรมนิยม (behaviourism) ในช่วงทศวรรษที่ 1940s -1950s ไปสู่การศึกษาภาษาตามแนวทางความเชื่อแบบเหตุผลนิยม (rationalism)) โดยที่ในยุคก่อน Chomsky นั้น นักภาษาศาสตร์มองกระบวนการของการเรียนรู้ภาษาของมนุษย์ว่าเป็นไปในลักษณะเดียวกับการตอบสนองต่อสิ่งเร้า คนเราเรียนรู้

ภาษาจากการจำ จากการใช้ จากการที่ได้พบเห็นในชีวิตประจำวัน ซึ่งหากเชื่อเช่นนี้ ภาษาก็มีลักษณะของการเป็นสิ่งที่มนุษย์ประดิษฐ์ขึ้น ซึ่งในสังคมที่ต่างกลุ่มกัน ก็มีการใช้ภาษาที่แตกต่างกัน การใช้ภาษาใดได้อย่างถูกต้องเป็นสิ่งที่ต้องเรียนรู้จากสังคมนั้นๆ ซึ่งเมื่อพิจารณาแล้วแนวคิดแบบนี้ก็น่าจะเหมาะสม เพราะแต่ละสังคมก็มีภาษาที่ต่างกัน คนจีนพูดภาษาจีน คนไทยพูดภาษาไทย แต่ถ้าหากเราพิจารณาดูผลิตผลอื่นๆทางวัฒนธรรม เช่น เรื่องของเทคโนโลยี เราจะพบว่าเทคโนโลยีในสังคมต่างๆ นั้นมีความแตกต่างกันเป็นอันมาก และเทคโนโลยีในยุคต่างๆ ก็มีความแตกต่างอย่างเห็นได้ชัด กล่าวคือ เรามีเทคโนโลยีในยุคหิน เทคโนโลยีในยุคโลหะ ต่อเนื่องมาจนถึงเทคโนโลยีในยุคคอมพิวเตอร์ ซึ่งจะเห็นว่ามีการเปลี่ยนแปลงในทางที่มีความสลับซับซ้อนมากขึ้น แต่เมื่อเราพิจารณาในเรื่องของภาษา เราจะไม่พบว่ามีความแตกต่างอย่างเด่นชัดในเรื่องของความซับซ้อนของภาษาระหว่างภาษาในยุคโบราณและภาษาในยุคปัจจุบัน ถึงแม้ภาษาต่างๆจะดูเสมือนต่างกันมาก แต่เมื่อพิจารณาในรายละเอียดของหลักการแล้ว ความแตกต่างทางภาษานั้นไม่ได้เป็นความแตกต่างในลักษณะที่ไร้ข้อจำกัด ตรงกันข้าม เรากลับพบว่ามีลักษณะที่สอดคล้องกันระหว่างภาษาต่างๆ เป็นอย่างมาก

ถ้าเช่นนั้น อะไรที่ทำให้ Chomsky เชื่อว่าความสามารถทางภาษาเป็นสิ่งที่มนุษย์มีมาโดยกำเนิดของหรือที่ Chomsky เรียกว่า innate การที่ภาษามีลักษณะเป็นสากล (universal) คือไม่มีสังคมใดที่ไม่มีภาษาจะถือว่าเป็นเครื่องบ่งชี้ความเป็น innate ได้หรือไม่ Pinker (1994) กล่าวถึงเรื่องนี้ไว้ในหนังสือ The language Instinct ว่า การที่จะกล่าวว่าไม่มีสังคมใดเลยในโลกนี้ที่มนุษย์จะอยู่อย่างไร้ภาษา แล้วมาสรุปว่าภาษาเป็นสิ่งที่มีมาโดยธรรมชาตินั้นไม่ใช่เหตุผลที่ดีพอ ตัวอย่างเช่น โคคาโคลาอาจจะไปปรากฏอยู่ในทุกสังคมมนุษย์ที่มีอยู่ในโลกนี้ แต่ก็ไม่ได้หมายความว่า โคคาโคลานั้นเป็นสิ่งที่มีมาเองโดยธรรมชาติ ดังนั้น การที่มนุษย์ทุกสังคมมีภาษานั้นจึงไม่ได้เป็นเครื่องยืนยันว่าภาษาเป็นสิ่งที่มีมาเองโดยธรรมชาติเพราะภาษาก็อาจเป็นสิ่งที่ประดิษฐ์ขึ้นเหมือนกับโคคาโคลาก็ได้

แต่ Pinker กล่าวว่า เหตุผลที่ทำให้เชื่อได้ว่าภาษาเป็นสิ่งที่มีมาโดยธรรมชาติหรือความสามาถทางภาษาของมนุษย์เป็นสิ่งที่มีติดตัวมาตั้งแต่เกิด ได้มาจากการศึกษาเรื่องการเรียนรู้ภาษาของเด็ก (language acquisition) ตัวอย่างเช่น ในยุดก่อน ในสมัยที่มีการค้าทาส ได้มีการจับคนจากเผ่าต่างๆลงในใต้ท้องเรือลำเดียวกัน คนเหล่านี้มาจากต่างเผ่า พูดภาษาที่ต่างกัน แต่คนเหล่านี้ก็มีความพยายามที่จะสื่อสารกัน โดยพยายามที่จะเรียนรู้ภาษาของอีกฝ่ายหนึ่ง แต่ก็ไม่สามารถใช้ภาษาของอีกเผ่าหนึ่งได้สมบูรณ์ เป็นภาษาที่มีลักษณะครึ่งๆกลางๆ ไม่มีกฎเกณฑ์ทางไวยากรณ์ที่เป็นระบบ หรือที่เรียกว่า ภาษา pidgin ตัวอย่างเช่น ภาษาอังกฤษของผู้อพยพจะมีลักษณะเป็นคำเป็นวลีที่ผิดกฎไวยากรณ์ของภาษาอังกฤษ เช่น พูดว่า me no hungry แทนที่จะพูดว่า I'm not hungry สิ่งที่น่าสนใจก็คือ ในกลุ่มคนที่พูดภาษา pidgin นี้ เมื่อมีลูกหลานและเด็กรุ่นลูกของคนกลุ่มนี้เมื่อมาอยู่รวมกันเป็นกลุ่ม เด็กเหล่านี้จะสามารถพัฒนาระบบภาษาใหม่ขึ้นจากภาษา pidgin ที่ได้ยินจากพ่อแม่ เกิดเป็นภาษาที่มีระบบไวยากรณ์สมบูรณ์ในตัวเองที่เรียกว่าภาษา creole ได้ สิ่งเหล่านี้แสดงให้เห็นว่า เด็กเหล่านี้ไม่ได้เรียนภาษาจากการเลียนแบบหรือจากการได้ยินได้ฟังในสังคมเหมือนอย่างการเรียนรู้ทางวัฒนธรรมอื่นๆ แต่เป็นการใช้ความสามารถทางภาษาที่เด็กเหล่านี้มีอยู่ในตัวแต่กำเนิดมาพัฒนาจัดระบบของข้อมูลภาษาแบบ pidgin ที่ได้ยินให้เป็นระบบภาษาที่สมบูรณ์มีกฎเกณฑ์ทางไวยากรณ์ของตัวเองขึ้นมาได้

การเรียนรู้ภาษานี้เป็นคุณสมบัติที่พบเฉพาะในเด็กเล็ก เมื่อผ่านพ้นช่วงวัยหนึ่งไปแล้ว การเรียนรู้ภาษาโดยธรรมชาติก็จะเป็นเรื่องยาก ในเรื่องนี้ มีความเชื่อทั่วๆไปอยู่ว่า การที่แม่พยายามพูดภาษากับเด็กอยู่เสมอๆโดยใช้รูป

ประโยคแบบง่ายๆ จะมีผลต่อพัฒนาการทางภาษาที่ดีของเด็ก ภาษาที่แม่พูดกับเด็กโดยเฉพาะนี้เรียกว่า motherlease แต่ก็มีผู้แย้งว่าเด็กสามารถเรียนรู้ภาษาได้เองจากบริบทที่มีการใช้ภาษาเกิดขึ้น ไม่จำเป็นต้องอาศัยภาษาในแบบ motherlease นี้ Chomsky กล่าวว่าเด็กไม่ได้เรียนรู้ภาษาจากการจดจำประโยคที่ได้ยิน โดยยกตัวอย่างการสร้างประโยคคำถามในภาษาอังกฤษว่าเด็กสามารถสร้างประโยคคำถามได้ถูกต้องแม้จะไม่เคยได้ยินประโยคแบบนั้นมาก่อน ตัวอย่างเช่น ประโยคอย่าง A unicorn that is eating a flower is in the garden ซึ่งเป็นประโยคแบบที่เด็กจะไม่เคยได้ยินได้ฟังเพราะมีโครงสร้างซับซ้อนเกินกว่าที่จะพบในภาษา motherlease แต่เมื่อทดลองให้เด็กสร้างประโยคคำถามจากประโยคนี้ กลับไม่มีเด็กคนไหนสร้างประโยคคำถามผิดๆเป็น Is a unicorn that eating a flower is in the garden? ซึ่งเกิดจากการย้าย is ตัวแรก แต่เด็กกลับสร้างประโยคคำถามได้ถูกต้องโดยเลือกย้าย is ตัวที่สองเป็น Is a unicorn that is eating a flower in the garden? ซึ่งหากเราเชื่อว่าเด็กเรียนรู้ภาษาจากการเลียนแบบ เด็กก็น่าจะเลือกย้าย is ตัวแรกมากกว่า เพราะประโยคที่เด็กจะได้ยินได้ฟังอยู่เสมอคือประโยคความเดียวอย่าง A unicorn is in the garden และประโยคคำถามอย่าง Is a unicorn in the garden? ทำไมเด็กจึงสร้างประโยคคำถามในประโยคซับซ้อนได้ถูกต้อง ทั้งๆที่ประโยคซับซ้อนแบบนี้คงจะไม่พบในภาษา motherlease หรือในประโยคที่แม่พูดกับลูก ทำไมเด็กจึงไม่ใช้กฎง่ายๆที่ว่าให้ย้ายกริยาช่วยตัวแรกที่พบ การที่เด็กเรียนรู้ภาษาได้จากข้อมูลอินพุทที่ไม่ครบบริบูรณ์ (poverty of input) นี้เองที่ทำให้ Chomsky เชื่อว่าความสามารถทางภาษาเป็นคุณสมบัติที่มนุษย์ทุกคนมี

นอกจากนี้ กรณีของผู้ป่วยที่ได้รับอุบัติเหตุทางสมองแล้วสูญเสียความสามารถทางภาษาไป ก็ทำให้เราเห็นว่ามีสมองบางส่วนที่ทำหน้าที่เฉพาะทางด้านภาษาแยกจากการทำหน้าที่คิดหรือรับรู้ในเรื่องอื่นๆ เช่น ในกรณีผู้ป่วยที่เป็นโรค Broca's aphasia ผู้ป่วยจะสูญเสียความสามารถทางไวยากรณ์ (grammatical impair) คือ จะไม่สามารถใช้ประโยคที่ถูกตามไวยากรณ์เหมือนอย่างที่เคยใช้มาแต่ก่อนได้ แต่ความสามารถในด้านอื่นๆของผู้ป่วยยังคงเป็นปกติดี นอกจากนี้ก็ยังมีโรคที่ทำให้ผู้ป่วยมีความสามารถทางภาษามากเกินปกติ (language fluent) เช่น ในผู้ที่เป็น autistic บางคนจะมีความสามารถพิเศษทางภาษาสามารถใช้ศัพท์ยากๆสำนวนหรูๆที่คนปกติไม่ใช้ ในขณะที่ความสามารถทางสติปัญญาด้านอื่นกลับสู้คนปกติไม่ได้ สิ่งเหล่านี้ชี้ให้เห็นว่า ความสามารถทางภาษานี้เป็นส่วนที่แยกต่างหากจากความสามารถด้านอื่นๆ เช่น การคิด การใช้เหตุผล การมองเห็น การได้ยิน เป็นต้น

กล่าวโดยสรุป นักภาษาศาสตร์ในกลุ่มของ Chomsky เชื่อว่าความสามารถทางภาษาของมนุษย์เป็นคุณสมบัติที่มนุษย์ทุกคนมีมาแต่กำเนิด เมื่อเด็กได้อยู่ในบริบทสังคมที่ใช้ภาษาใด ข้อมูลภาษาที่ได้ยินจะไปกระตุ้นให้สมองของเด็กพัฒนาระบบของภาษานั้น ๆ ขึ้นมา คำถามที่สำคัญต่อไปก็คือ ระบบภาษาที่มนุษย์มีอยู่ในแต่ละภาษานั้นมีลักษณะอย่างไร มีความเหมือนหรือแตกต่างกันอย่างไร คำถามเหล่านี้เป็นคำถามที่นักภาษาศาสตร์ให้ความสนใจศึกษามาตั้งแต่ศตวรรษที่ 19 โดยบุคคลแรกที่เป็นผู้บุกเบิกการศึกษาภาษาศาสตร์อย่างเป็นระบบก่อนหน้า Chomsky และถือเป็นบิดาแห่งสาขาภาษาศาสตร์สมัยใหม่ คือ Ferdinand De Saussure

<u>ศาสตร์แห่งภาษา</u>

Saussure เป็นชาวสวิส เกิดในปี 1857 หนึ่งปีหลังจาก Sigmund Freud บิดาแห่งจิตวิทยาสมัยใหม่ และหนึ่งปีก่อน Emile Durkheim บิดาแห่งสังคมวิทยา Saussure ถือเป็นบิดาแห่งภาษาศาสตร์สมัยใหม่ โดย

Saussure เตือนให้นักภาษาศาสตร์เห็นว่าการศึกษาภาษาในลักษณะที่เน้นการศึกษาเรื่องประวัติความเป็นมาของคำและวิวัฒนาการของภาษานั้นไม่สามารถทำให้เราเข้าใจธรรมชาติที่แท้จริงของภาษาได้

Saussure ไม่ได้เขียนและตีพิมพ์หนังสือรวบรวมความรู้ของตัวเอง แต่บรรดาลูกศิษย์ของเขาได้รวบรวมสมุดโน้ตของลูกศิษย์หลายๆ รุ่นซึ่งจดจากการฟังการบรรยายของเขามารวมและตีพิมพ์เป็นหนังสือในปี 1915 ชื่อ Cours de linguistique generale (A course in General Linguistics) Saussure มีความคิดว่า นักภาษาศาสตร์ขณะนั้นยังหลงประเด็น และสับสนกับงานของตัวเอง เขาจึงมองเห็นความจำเป็นที่จะต้องปฏิรูปศาสตร์นี้

ในยุคก่อนหน้า Saussure นี้ การศึกษาภาษาได้ผ่านการเปลี่ยนแปลงใหญ่ๆมาแล้วสามยุค โดยในยุคแรกเป็นยุคที่มองภาษาตามแบบไวยากรณ์ดั้งเดิม คือสนใจในเรื่องของกฎเกณฑ์การใช้ภาษา เรื่องของความถูกผิดในการใช้ โดยมีเกณฑ์เรื่องความสละสลวยของภาษาเขียนที่ดีเป็นเกณฑ์หลัก ยุคที่สองเป็นยุคของนิรุกติศาสตร์ สนใจศึกษาภาษาเก่าๆ ศึกษาความเป็นมาของคำแต่ละคำ และยุคที่สามเป็นยุคของการเปรียบเทียบภาษา หาความสัมพันธ์ระหว่างภาษาต่างๆ ซึ่งที่ผ่านมาไม่ได้ทำให้เกิดการพัฒนาภาษาศาสตร์อย่างเป็นวิทยาศาสตร์อย่างแท้จริง เพราะไม่สามารถอธิบายถึงธรรมชาติที่แท้จริงของภาษาได้

สิ่งที่ Saussure เสนอคือให้มีการศึกษาเรื่องของสัญลักษณ์หรือศาสตร์ด้านสัญญะวิทยา (semiology) ขึ้นมา โดยที่ภาษาศาสตร์เป็นส่วนหนึ่งของสัญญะวิทยานี้ Saussure ถือว่าภาษาเป็นระบบสัญลักษณ์ที่มีความซับซ้อนมากที่สุด ควรที่จะเป็นต้นแบบของการศึกษาระบบสัญลักษณ์อื่นๆ สาเหตุที่ไม่มีใครมองเห็นอย่างที่ Saussure เห็นมาก่อน อาจเป็นเพราะว่าแต่ก่อนนั้นการศึกษาภาษาเป็นการศึกษาภาษาในลักษณะที่ผูกติดกับสิ่งอื่นๆ คือผูกภาษาอยู่กับความงดงามของภาษา ผูกภาษาอยู่กับประวัติศาสตร์ ผูกภาษาอยู่กับวิวัฒนาการ Saussure จึงประกาศว่าวัตถุประสงค์ที่แท้จริงของภาษาศาสตร์ควรจะเป็นการศึกษาภาษาเพื่อความเข้าใจในธรรมชาติของภาษานั้นเอง

แม้ว่าภาษาจะไม่ได้มีลักษณะทางกายภาพที่จับต้องได้โดยตรงเหมือนวัตถุต่างๆ แต่ Saussure ก็มองว่าภาษาเป็นสิ่งที่มีตัวตนที่สามารถศึกษาได้ ซึ่งสิ่งนั้นเป็นผลผลิตร่วมทางสังคมที่คนเราทุกคนมีอยู่ในหัว สิ่งนั้นก็คือสัญลักษณ์ ซึ่งเกิดจากการรวมกันของรูปเสียงและมโนทัศน์ บางคนอาจจะคิดว่าภาษาเป็นเรื่องของการตั้งชื่อ (namimg) สิ่งต่างๆในโลก คำแต่ละคำมีขึ้นเพื่อใช้เรียกสิ่งที่มันแทน ความคิดแบบนี้ตั้งอยู่บนข้อสมมุติที่ว่ามีความคิดที่สำเร็จรูปอยู่ก่อนที่จะมีคำ ซึ่งหากเป็นเช่นนั้น ความแตกต่างระหว่างภาษาทั้งหลายในโลกก็น่าจะเป็นเพียงการเรียกชื่อสิ่งต่างๆแตกต่างกัน แต่ในความเป็นจริง คำซึ่งแทนความหมายอย่างหนึ่งในภาษาอาจจะไม่มีคำที่มีความหมายเหมือนกันในอีกภาษาหนึ่ง ดังนั้นสัญลักษณ์ในภาษาจึงไม่ใช่เป็นเพียงการตั้งชื่อให้กับสิ่งของในโลกและคนในแต่ละสังคมก็ไม่ได้มีความคิดสำเร็จรูปที่เหมือนกัน

สัญลักษณ์ในภาษามีลักษณะที่เป็นหน่วยทางความคิด (psychological entity) ถ้าเราลองให้แต่ละคนพูดว่า “แมวนอนอยู่บนเสื่อ” แล้วใช้เครื่องบันทึกเสียงเพื่อดูภาพคลื่นเสียงที่ได้ เราจะเห็นว่าคลื่นเสียงที่ได้จากแต่ละคนนั้นแตกต่างกัน แต่ทำไมคนเราถึงสามารถสื่อสารกันได้ ทำไมคนเราจึงสามารถเข้าใจประโยคนี้ได้ตรงกัน ก็เพราะว่าทุกคนเข้าใจสัญลักษณ์ต่างๆ ที่ใช้ในประโยคนี้ได้ตรงกัน สัญลักษณ์จึงเป็นหน่วยที่เป็นนามธรรมไม่ใช่ตัวคลื่นเสียงที่จับมาวัดได้

สัญลักษณ์ประกอบไปด้วยส่วนที่เป็น signifier คือส่วนที่เป็นสื่อหรือพาหะที่ใช้แทนสัญลักษณ์นั้นและส่วนที่เป็น signified หรือคือส่วนที่เป็นความหมายที่คนในสังคมมีร่วมกันสำหรับสัญลักษณ์นั้น ในทางภาษาแล้ว signifier คือลำดับของหน่วยเสียงต่างๆในคำนั้น และ signified คือมโนทัศน์ที่คำนั้นสื่อถึง ความสัมพันธ์ระหว่าง signifier กับ signified นี้มีลักษณะที่ไม่มีกฎเกณฑ์ตายตัว (arbitrary) คือไม่ได้ถูกกำหนดไว้ก่อน เช่น มโนทัศน์ของคำว่า “horse” ไม่ได้ขึ้นอยู่กับรูปเสียง /hos/ ที่ใช้ ในภาษาอื่นๆก็จะใช้รูปเสียงอื่นๆ ที่ต่างกัน เช่น ภาษาไทยใช้รูปเสียง /maa2/ ความไม่มีกฎเกณฑ์นี้ปรากฏในระบบสัญลักษณ์อื่น ๆ ด้วย เช่น การแสดงความนอบน้อมในแต่ละสังคมก็มีวิธีการที่ต่างกัน แต่การที่บอกว่าสัญลักษณ์ในภาษามีลักษณะที่ไม่มีกฎเกณฑ์นั้นไม่ได้หมายความว่า ใครอยากจะกำหนดใช้รูปเสียงอะไรแทนมโนทัศน์ใดก็ได้ เพราะสัญลักษณ์ไม่ใช่เรื่องของคนใดใคนหนึ่งแต่จะต้องเป็นสิ่งที่เกิดจากความเข้าใจร่วมกันของคนในสังคมนั้นเอง

การกำหนดสัญลักษณ์ต่างๆขึ้นมาใช้ทำให้เราสามารถแยกความแตกต่างระหว่างความคิดต่าง ๆ ได้ หากไม่มีการกำหนดใช้สัญลักษณ์ ความคิดต่าง ๆ ที่คนเรามีก็เป็นเสมือนบางสิ่งบางอย่างที่ยังไม่มีการแบ่งแยกออกเป็นส่วนต่าง ๆชัดเจน การมีสัญลักษณ์ต่างๆทำให้มีการแบ่งแยกความคิดออกเป็นส่วนๆ ได้เรียกว่ามโนทัศน์หรือ concept มีการจัดระเบียบความคิดโดยการเชื่อมโยงความสัมพันธ์ระหว่างแต่ละมโนทัศน์กับรูปเสียงที่ใช้แทน ภาษาจึงเปรียบเสมือนเหรียญที่มี 2 ด้าน ด้านหนึ่งแทนความคิดหรือมโนทัศน์อีกด้านหนึ่งแทนรูปเสียงที่ใช้ เมื่อรวมทั้งสองด้านเข้าด้วยกันก็จะเกิดสิ่งเป็นสัญลักษณ์ขึ้น ณ จุดนี้ เราได้ภาพความเข้าใจว่าภาษาประกอบด้วยสัญลักษณ์ต่าง ๆ ที่คนในสังคมกำหนดใช้ร่วมกัน แต่การมองเพียงแค่นี้ก็ยังไม่เพียงพอ เราไม่สามารถมองสัญลักษณ์โดยอิสระจากระบบได้ แต่เราจะต้องพิจารณาความสัมพันธ์ระหว่างสัญลักษณ์ที่อยู่ภายในระบบด้วย

Saussure กล่าวว่าเราไม่สามารถศึกษาสัญลักษณ์โดยไม่สนใจระบบของสัญลักษณ์ที่เป็นอยู่ได้ เพราะคุณค่า (value) ของสัญลักษณ์แต่ละตัวไม่ได้อยู่ที่ตัวสัญลักษณ์นั้น แต่คุณค่าที่แท้จริงอยู่ที่การมีอยู่ของสัญลักษณ์อื่นๆในระบบที่ทำให้สัญลักษณ์ตัวนี้ต่างจากสัญลักษณ์ตัวอื่น ๆ ตัวอย่างเช่น เมื่อพิจารณาจากด้านมโนทัศน์หรือความหมายของคำ คำว่า mouton ในภาษาฝรั่งเศสอาจใช้เพื่อสื่อถึงมโนทัศน์เดียวกับคำว่า sheep ในภาษาอังกฤษ แต่สัญลักษณ์ทั้งสองอันนี้มีคุณค่าไม่เท่ากัน เพราะว่าในภาษาอังกฤษมีคำว่า moutton ที่หมายถึงเนื้อที่พร้อมเสริฟบนโต๊ะอาหารด้วย ในขณะที่ภาษาฝรั่งเศสยังใช้คำเดิมคือ mouton หรือเมื่อพิจารณาจากด้านรูปเสียง คุณค่าของสัญลักษณ์ก็เป็นไปในลักษณะเดียวกันคือขึ้นอยู่กับความสัมพันธ์และความแตกต่างระหว่างรูปเสียงในระบบนั้นเป็นหลัก กล่าวคือ เรามองความต่างของฟอร์ม (form) เป็นสำคัญ โดยที่ฟอร์มหมายถึงหน่วยเสียง (phoneme) ที่คนในภาษานั้นรับรู้ เราไม่มองที่ความต่างของกายภาพ (substance) หรือความต่างของคลื่นเสียงที่เราได้ยิน ดังนั้นหน่วยในทางภาษาจึงไม่สามารถระบุ (identify) ได้จากคุณสมบัติในตัวเอง แต่ถูกระบุได้จากความแตกต่างจากหน่วยอื่นๆในภาษา คำอธิบายนี้ใช้ได้กับทั้งภาษาพูดและภาษาเขียน ตัวอย่างเช่น ตัวอักษร t อาจจะเขียนในรูปแบบต่างๆกัน ดังนี้ t t t t t t t แต่การที่เราสามารถรับรู้ได้ว่ามันเป็นตัวอักษร t เป็นเพราะเราเห็นความแตกต่างของตัวอักษรนี้จากอักษรตัวอื่นๆที่ไม่ใช่ t

นอกจากนี้ Saussure ยังชี้ให้เห็นถึงความแตกต่างของระบบภาษา (langue) กับการใช้ภาษา (parole) โดยที่ langue หมายถึงภาษาที่เป็นระบบของสัญลักษณ์ที่คนในสังคมเข้าใจร่วมกัน ส่วน parole หมายถึงภาษาที่เป็นคำพูดที่คนเราพูดออกมา คำพูดหรือประโยคทั้งหลายนี้เป็นสิ่งเดียวที่เราได้ยินได้ฟังที่สามารถนำมาศึกษาเพื่อ

หา langue หรือตัวระบบของภาษานั่นเอง Saussure แยกความต่างระหว่าง langue กับ parole เพื่อชี้ให้เห็นชัดว่าอะไรคือขอบเขตที่นักภาษาศาสตร์ควรศึกษา ซึ่งการจะศึกษาหา langue นั้นคือการหาระบบของภาษาซึ่งมีอยู่ ณ ช่วงเวลานั้น ไม่ใช่การศึกษาภาษาในลักษณะเปรียบเทียบตามช่วงเวลาต่างๆอย่างที่นักภาษาศาสตร์ได้ทำกันมา Saussure เปรียบให้เห็นโดยเทียบกับเกมหมากรุก โดยให้นึกถึงสภาวะของเกมหมากรุกที่ช่วงเวลาใดเวลาหนึ่ง เราสามารถอธิบายว่าหมากตัวใดวางอยู่ณตำแหน่งใดและสามารถเดินทางใดได้บ้าง ซึ่งข้อมูลว่าหมากตัวนั้นเคยเดินมาจากที่ไหนบ้างนั้นไม่สำคัญต่อการอธิบายสภาวะของเกมส์ในขณะนั้น ภาษาก็เหมือนกัน ถึงแม้ว่าจะมีการเปลี่ยนแปลงของภาษาเกิดขึ้นตามช่วงเวลาต่างๆ แต่ ณ ขณะนั้น ความเปลี่ยนแปลงที่เคยเกิดขึ้นมาไม่ได้มีความสำคัญต่อการรับรู้ภาษาของคนในสังคม ณ เวลานั้น เราไม่จำเป็นต้องรู้ว่าคำๆนี้เมื่อก่อนเคยมีความหมายอะไรมาบ้าง จึงจะสามารถใช้คำๆนั้นได้ถูกต้อง ขอเพียงแค่รู้ว่าคำๆนี้มีความหมายในปัจจุบันอย่างไรก็เพียงพอแล้ว

นอกจากนี้ Saussure ยังให้ความสำคัญกับภาษาพูดด้วย โดยชี้ให้เห็นถึงความสำคัญของภาษาพูดว่าภาษาพูดนั้นมีมานานก่อนภาษาเขียน ภาษาเขียนโดยทั่วไปนั้นถูกสร้างขึ้นมาโดยอาศัยหน่วยทางเสียงเป็นพื้นฐาน[1] ความแตกต่างอีกประการของภาษาเขียนและภาษาพูด คือ ตัวเขียนไม่สามารถแสดงระดับสูงต่ำของเสียง (pitch) หรือแสดงการเน้นลงเสียงหนัก (stress) ในภาษาพูดได้ การใช้ตัวอักษรเอียงตัวอักษรเข้มเป็นเพียงวิธีทางอ้อมที่เราพยายามใช้เพื่อเน้นให้เห็นความแตกต่างจากส่วนอื่นๆ และในบางภาษาความแตกต่างระหว่างภาษาเขียนและภาษาพูดก็เห็นชัดเจน เช่น ในประเทศจีนมีภาษาพูดแบบต่างๆ มากมาย แต่คนจีนสามารถใช้ระบบตัวเขียนเดียวกันเพื่อติดต่อสื่อสารกันด้วยตัวเขียนได้ถึงแม้ว่าจะฟังภาษาของอีกฝ่ายไม่เข้าใจ ลักษณะนี้เป็นกรณียกเว้น ที่ภาษาเขียนได้พัฒนาแยกจากภาษาพูดอย่างชัดเจน อีกตัวอย่างคือ ภาษาลาตินและภาษาสันสกฤตที่เป็นภาษาที่ตายไปแล้วคือไม่มีสังคมใดที่พูดภาษานี้เป็นภาษาแม่ จึงไม่มีการเรียนรู้โดยกำเนิดเหมือนภาษาอื่นๆ แต่ต้องเรียนจากตัวเขียนอย่างเดียว ภาษาเขียนและภาษาพูดจึงเป็นระบบสัญลักษณ์ที่แยกจากกันเป็นสองระบบ โดยที่ภาษาเขียนมีขึ้นเพื่อใช้แทนภาษาพูดอีกที

<u>ระบบไวยากรณ์</u>

ณ จุดนี้ เรารู้ว่าภาษาเป็นระบบของสัญลักษณ์ โดยที่คำๆหนึ่งเป็นสัญลักษณ์ที่ใช้แทนมโนทัศน์หนึ่งๆ และคุณค่าของสัญลักษณ์แต่ละตัวอยู่ที่ความแตกต่างจากสัญลักษณ์อื่นๆในระบบ แต่ระบบของภาษานั้นซับซ้อนมากกว่าการเป็นแหล่งที่รวมของสัญลักษณ์จำนวนมากมาย กล่าวคือ ถึงแม้เราจะสามารถจำคำทั้งหมดในภาษาได้ว่ามีความหมายอย่างไร แต่ความรู้เท่านี้ยังไม่เพียงพอที่จะทำให้เราเข้าใจภาษานั้นได้ เช่น เมื่อเราต้องการกล่าวถึงสุนัขกับแมวและเหตุการณ์ที่สุนัขเป็นผู้กัดแมว เราไม่สามารถนำคำสามคำในภาษาไทยคือ “สุนัข” “แมว” และ “กัด” มาเรียงกันตามใจชอบได้เป็น “สุนัขแมวกัด” หรือ “กัดสุนัขแมว” หรือ “แมวกัดสุนัข” ได้ เพราะสิ่งที่ได้อาจไม่ใช่ประโยคที่สื่อความได้หรืออาจเป็นประโยคที่ไม่ได้สื่อความหมายอย่างที่ต้องการ ดังนั้น ในระบบของภาษาจึงมีสิ่งที่

---

[1] ถึงแม้ภาษาเขียนจะพัฒนาบนรากฐานของภาษาพูด แต่ผลจากการเปลี่ยนแปลงในภาษา ทำให้เกิดคำที่เขียนต่างกันแต่ออกเสียงเหมือนกัน เรียกว่า homophones และทำให้เกิดคำที่เขียนเหมือนกันแต่ออกเสียงต่างกัน เรียกว่า homograph เช่น lead, read ยิ่งภาษาเขียนเกิดมานานเท่าใด ความไม่สอดคล้องระหว่างตัวเขียนกับเสียงก็ยิ่งมีมากขึ้นตามไปด้วย

เรียกว่าระบบไวยากรณ์ (grammatical system) หรือกฎเกณฑ์เฉพาะของภาษานั้นๆ ซึ่งเมื่อเราพิจารณาการนำคำมารวมกันเป็นประโยคนี้ เราจะเห็นความสัมพันธ์ระหว่างคำในภาษาอยู่สองลักษณะตามที่ Saussure กล่าวถึง คือ ความสัมพันธ์ในแนวราบ (syntacmatic relationship) และความสัมพันธ์ในแนวดิ่ง (paradigmatic relationship) ความสัมพันธ์ในแนวราบเป็นความสัมพันธ์ที่เกิดขึ้นจากการนำหน่วยทางภาษามาจัดเรียงต่อกัน ซึ่งเราได้เห็นแล้วว่าหน่วยบางหน่วยไม่สามารถนำมาเรียงต่อกันแล้วเป็นหน่วยทางภาษาที่ยอมรับได้ เช่น "สุนัขแมวกัด" ทั้งนี้เพราะในภาษาไทยมีกฎทางไวยากรณ์ที่กำหนดไว้ว่าประธานจะเกิดหน้ากริยาและกรรมเกิดอยู่หลังคำกริยา ดังนั้น ประโยคที่ถูกคือ "สุนัขกัดแมว" ส่วนความสัมพันธ์ในแนวดิ่งหมายถึงการที่หน่วยทางภาษาสามารถจัดเป็นกลุ่มๆได้ โดยที่สมาชิกของหน่วยทางภาษาในแต่ละกลุ่มนั้นจะมีคุณสมบัติบางอย่างร่วมกัน เช่น ในประโยค "สุนัขกัดแมว" นี้ เราสามารถใช้คำอื่นแทนคำว่า "สุนัข" ได้ เช่น "งูกัดแมว" "แมวกัดแมว" "เสือกัดแมว" เป็นต้น ดังนั้น คำว่า "สุนัข" "งู" "แมว" และ "เสือ" มีความสัมพันธ์ในแนวดิ่งคือเป็นคำที่จัดอยู่ในกลุ่มเดียวกันได้คือกลุ่มที่เราเรียกว่าคำนาม

ในที่นี้ จะเห็นว่าเราใช้คำว่า "หน่วยทางภาษา" ในการอธิบายความสัมพันธ์ทางแนวนอนและความสัมพันธ์ในแนวดิ่งแทนที่จะใช้คำว่า "คำในภาษา" ที่เป็นเช่นนี้เพราะความสัมพันธ์ทั้งสองแบบนี้สามารถปรากฏได้ในระดับที่เล็กกว่าคำและในระดับที่ใหญ่กว่าคำด้วย ในภาษาอังกฤษจะเห็นความสัมพันธ์ในหน่วยที่เล็กกว่าคำชัดเจนกว่าภาษาไทย เช่น คำว่า walking sleeping หรือ eating นั้น ประกอบด้วยสองส่วนคือส่วนที่เป็นรากคำและส่วนที่เป็นปัจจัย (suffix) "–ing" หน่วยที่เล็กกว่าคำในตัวอย่างนี้เราเรียกว่า หน่วยคำ (morpheme) ส่วนความสัมพันธ์ในหน่วยที่ใหญ่กว่าคำ เช่น ความสัมพันธ์ระหว่างวลีต่างๆในประโยค เช่น "a dog", "a big dog", "a very big dog" " a smart jumping dog of mine" เป็นการรวมกันของคำเป็นนามวลี และนามวลีนี้สามารถเกิดร่วมกับกริยาวลี "bites a cat" ได้ เป็นต้น

ดังนั้น ในการศึกษาระบบของภาษานั้น นักภาษาศาสตร์ต้องศึกษามากกว่าเรื่องของคำต่างๆ นักภาษาศาสตร์ต้องศึกษาตั้งแต่ในระดับที่เล็กที่สุดคือในระดับของเสียงเพื่อหาว่าในภาษานั้นๆ มีหน่วยเสียง (phoneme) อะไรบ้าง คำว่าหน่วยเสียงหมายถึงรูปเสียงที่คนในภาษานั้นรับรู้ว่ามีความแตกต่างจากรูปเสียงอื่นๆ ถือเป็นหน่วยทางภาษา ตัวอย่างเช่น ในภาษาไทย เรารับรู้ว่าเสียงของ ก ต่างจากเสียง ค ดังจะเห็นว่าคำว่า กา กับคำว่า คา นั้นเป็นคำที่ต่างกัน ดังนั้นคนไทยจึงรับรู้ว่าสองเสียงนี้เป็นคนละรูปเสียงกัน แต่ในภาษาอังกฤษ ผู้พูดภาษาอังกฤษเป็นภาษาแม่จะรับรู้เสียงสองเสียงนี้เป็นรูปเสียงเดียวกัน โดยผู้พูดภาษาอังกฤษจะออกเสียงตัวอักษร "k" เป็น ก เมื่ออยู่หลัง s แต่ออกเสียงเป็น ค ในกรณีอื่นๆ เช่น "sky" ออกเสียงเป็น สะกาย แต่ "king" ออกเสียงเป็น คิง ดังนั้นคนที่พูดภาษาอังกฤษโดยทั่วไปจึงมักมีปัญหาเวลาถูกสอนให้พูดคำไทยว่า "ไข่ไก่" ทั้งนี้เพราะเขาไม่รับรู้ถึงความแตกต่างของสองเสียงนี้

นอกจากการศึกษาเพื่อหาระบบของหน่วยเสียงในภาษาหรือที่เรียกว่าสัทวิทยา (phonology) แล้ว นักภาษาศาสตร์ยังต้องหาระบบของหน่วยที่ใหญ่กว่าหน่วยเสียงขึ้นมาคือการศึกษาในระดับที่เรียกว่าวจีวิภาค (morphology) ในระดับนี้ หน่วยที่ศึกษาคือหน่วยคำ (morpheme) ซึ่งเป็นหน่วยที่เกิดจากการนำหน่วยเสียงมารวมกันจนกระทั่งเป็นหน่วยที่สามารถสื่อความหมายได้ เช่น หน่วยเสียง ก กับหน่วยเสียงสระ อา เมื่อมารวมกันเกิดเป็นหน่วยคำขึ้นในภาษาไทยคือ "กา" /kaa/ ซึ่งหมายถึงสัตว์ปีกชนิดหนึ่ง ในภาษาไทยนี้จะไม่เห็นลักษณะของวจีวิภาคชัดเจนนัก เนื่องจากภาษาไทยไม่มีระบบทางวจีวิภาคที่ซับซ้อนเหมือนภาษาอื่น เช่น ภาษาอังกฤษ หรือ ภาษารัส

เซีย ซึ่งต้องมีการจัดกลุ่มหน่วยคำที่มีคุณสมบัติคล้ายกันเป็นกลุ่มและหาความสัมพันธ์ในแนวราบว่าหน่วยคำกลุ่มใดสามารถเกิดร่วมกับกลุ่มใดได้บ้าง สิ่งที่ได้จากการรวมกันของหน่วยคำคือคำ เช่น running เป็นคำที่มาจากสองหน่วยคำคือ หน่วยคำที่เป็นรากคำของกริยา run และหน่วยคำที่เป็นปัจจัย –ing

หลังจากนั้น จึงเป็นการศึกษาในระดับที่สูงกว่าคำ คือการศึกษาหาระบบหรือกฎเกณฑ์ในภาษาที่กำหนดลักษณะการรวมกันของคำเป็นวลีหรือประโยค หรือที่เราเรียกว่าวากยสัมพันธ์ (syntax) กฎเกณฑ์ทางวากยสัมพันธ์นี้ทำให้เราเข้าใจได้ว่า “สุนัขกัดแมว” นั้นมีความหมายแตกต่างจาก “แมวกัดสุนัข” ถึงแม้จะประกอบด้วยคำที่เหมือนกันสามคำ และยังบอกเราว่า “สุนัขแมวกัด” ไม่ใช่ประโยคที่คนไทยใช้กัน หน่วยที่ศึกษาในระดับนี้คือประโยค โดยนักภาษาศาสตร์พยายามหาระบบไวยากรณ์ของภาษานั้น ซึ่งจะช่วยอธิบายได้ว่าทำไมประโยคบางประโยคจึงผิดไวยากรณ์ ซึ่งความถูกต้องตามไวยากรณ์นี้เป็นคนละอย่างกับเรื่องการยอมรับได้ (acceptable) ของประโยคนั้นๆ เช่น Sam put a knife on the table เป็นประโยคที่ไม่ถูกตามไวยากรณ์เพราะไม่ได้เติม s หลังคำกริยา put แต่ประโยคนี้ถือเป็นประโยคที่ยอมรับได้และอาจพบเห็นบ้างในชีวิตจริง อาจจะเกิดจากการพูดหรือเขียนโดยไม่ระวังในขณะเดียวกัน ประโยคบางประโยคอาจถูกต้องตามไวยากรณ์แต่กลับเป็นประโยคที่ไม่สามารถยอมรับได้ เนื่องจากไม่มีผู้ใดใช้ เช่น The man the girl the boy knows likes is here. เป็นประโยคที่ยอมรับไม่ได้เพราะไม่มีใครใช้ แต่เป็นประโยคที่ถูกไวยากรณ์เพราะสามารถอธิบายได้ด้วยกฎชุดเดียวกับที่ใช้อธิบายประโยค [The man [the girl] likes]$_a$ is here. และประโยค [The girl the boy knows]$_b$ likes the man. ซึ่งเมื่อเทียบโครงสร้างประโยคแรกกับโครงสร้างของสองประโยคหลังแล้วจะได้ดังนี้ [The man [the girl the boy knows]$_b$ likes] $_a$ is here. (Pinker 1994) นักภาษาศาสตร์แต่ละกลุ่มได้พยายามพัฒนาทฤษฎีเพื่อนำมาใช้อธิบายระบบไวยากรณ์ของภาษา แนวคิดหนึ่งที่ใช้กันคือการมองว่าคำสามารถรวมกันเป็นวลีและวลีสามารถรวมกันเป็นวลีที่ใหญ่ขึ้นหรือเป็นประโยคได้ ซึ่งสามารถเขียนออกมาในรูปของกฎทางโครงสร้าง เช่น S -> NP VP, NP -> (DET) (ADJ) N, VP -> V NP เป็นต้น กฎ S -> NP VP บอกว่าประโยคประกอบด้วยสองส่วนส่วนแรกเป็นนามวลีส่องที่สองเป็นกริยาวลี กฎ NP -> (DET) (ADJ) N บอกว่านามวลีประกอบด้วยสามส่วน ส่วนแรกเป็น determiner ส่วนที่สองเป็นคำคุณศัพท์ ส่วนสุดท้ายเป็นคำนาม การใช้เครื่องหมายวงเล็บเพื่อบอกว่าส่วนนั้นๆสามารถละได้ กฎทางโครงสร้างลักษณะนี้สามารถใช้เพื่อแสดงโครงสร้างของประโยคเพื่อบอกว่าประโยคประกอบด้วยส่วนต่างๆอะไรบ้าง และลำดับชั้น (hierarchy) ขององค์ประกอบเหล่านั้นเป็นอย่างไร

นอกจากนี้ นักภาษาศาสตร์ยังสนใจมากกว่าการหาระบบไวยากรณ์ของภาษาใดภาษาหนึ่ง กล่าวคือนักภาษาศาสตร์ต้องการหาสิ่งที่เป็นระบบไวยากรณ์สากล (universal grammar) ซึ่งจะเป็นคำอธิบายได้ว่าทำไมมนุษย์จึงมีความสามารถทางภาษา ทำไมเด็กเล็กจึงสามารถเรียนรู้ที่จะพัฒนาระบบภาษาใด ๆ ก็ได้ที่ใช้กันอยู่ในสังคมนั้น ไวยากรณ์สากลนี้ตามความคิดของ Chomsky เป็นเสมือนกลไกทางภาษาที่มนุษย์ทุกคนมีเหมือนกันมาแต่กำเนิด แต่สาเหตุที่ลักษณะของภาษาแต่ละภาษานั้นแตกต่างกัน เป็นเพราะแต่ละภาษามีการตั้งค่าพารามิเตอร์ของระบบที่ต่างกัน เช่น บางภาษาเลือกใช้คำขยายอยู่ทางขวา บางภาษาเลือกให้คำขยายอยู่ทางซ้าย บางภาษาเลือกให้มีการละประธานได้ บางภาษาก็ไม่ยอมให้มีการละประธาน เป็นต้น ความแตกต่างของพารามิเตอร์เหล่านี้นี่เองที่เป็นเหตุให้ภาษาแต่ละภาษามีลักษณะที่ต่างกันไป ทำให้ระบบไวยากรณ์ของแต่ละภาษานั้นแตกต่างกัน

นอกจากการศึกษาในเรื่องของวากยสัมพันธ์แล้ว นักภาษาศาสตร์ยังสนใจในเรื่องของความหมายในภาษา กลไกที่ทำให้เราสามารถเข้าใจความหมายของประโยคแต่ละประโยค ซึ่งนอกจากความหมายที่ได้จากรูปประโยคโดยตรงที่ได้ยินแล้ว มนุษย์เรายังสามารถเข้าใจความหมายโดยอ้อมที่แฝงอยู่ในประโยคด้วย เช่น สามารถเข้าใจได้ว่า ทำไมประโยคแบบเดียวกันอย่าง "ใช่ คุณเก่งมาก" บางครั้งใช้ในทำนองเสียดสีแทนที่จะเป็นคำชมได้ ศาสตร์ที่ศึกษาเรื่องความหมายที่ได้จากการตีความรวมของความหมายย่อยๆในประโยคเรียกว่าอรรถศาสตร์ (semantics) มีแนวคิดต่างๆที่พยายามอธิบายความหมายของประโยค บ้างก็อธิบายในลักษณะที่เป็นการรวมกันของอรรถลักษณ์ (semantic feature) ต่างๆ เช่น ความหมายของคำว่า boy เกิดจากอรรถลักษณ์ +human +young +male เป็นต้น บ้างก็เห็นว่าจะต้องอธิบายโดยใช้ภาษาที่แน่ชัดไม่มีความกำกวมอย่างภาษาทางคณิตศาสตร์เช่น predicate logic เพื่อแทนความหมายของแต่ละประโยค ส่วนศาสตร์ที่ศึกษาความหมายที่เกิดจากการใช้ในบริบทจริงๆ เรียกว่า วัจนปฏิบัติศาสตร์ (pragmatics) ศาสตร์นี้เป็นศาสตร์ที่กว้างและไม่มีขอบเขตที่ชัดเจน ครอบคลุมการศึกษาการใช้ภาษาที่เกิดขึ้นในบริบท

นอกจากนี้ นักภาษาศาสตร์ยังต้องศึกษาภาษาในระดับที่ใหญ่กว่าประโยค คือระดับที่เป็นปริเฉจ (discourse) หรือระดับที่ประโยคมาเรียงต่อกันเป็นข้อความต่อเนื่อง ซึ่งต้องศึกษาเรื่องของการเชื่อมโยงความระหว่างประโยค การใช้สรรพนาม หรือนามชี้เฉพาะเพื่ออ้างถึงสิ่งที่เคยกล่าวอ้างมาแล้วในประโยคก่อนๆ เป็นต้น

นอกจากการศึกษาในตัวระบบของภาษาแล้ว นักภาษาศาสตร์บางกลุ่มก็สนใจศึกษาความสัมพันธ์ระหว่างภาษากับเรื่องอื่นๆ เช่น ภาษาศาสตร์สังคมศึกษาเรื่องของสังคมผ่านทางภาษา ภาษาศาสตร์กับการสอนภาษาศึกษาเรื่องการประยุกต์ใช้ความรู้ทางภาษาศาสตร์เพื่อประโยชน์ในการสอนภาษา และภาษาศาสตร์คอมพิวเตอร์ที่ศึกษาเรื่องของภาษาเพื่อนำมาประยุกต์ใช้กับคอมพิวเตอร์

<u>การพัฒนาระบบประมวลผลภาษาธรรมชาติตามแนวทางภาษาศาสตร์</u>

เมื่อได้เห็นภาพโดยคร่าวๆของการศึกษาทางภาษาศาสตร์แล้ว ซึ่งประกอบไปด้วยการศึกษาหน่วยภาษาในระดับต่างๆ ตั้งแต่ระดับสัทวิทยา ระดับวจีวิภาค ระดับวากยสัมพันธ์ ระดับอรรถศาสตร์ ระดับวัจนปฏิบัติศาสตร์ ไปจนถึงในระดับปริเฉจ ในส่วนนี้ เราจะพิจารณาถึงการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ตามความเข้าใจเรื่องของภาษาในระดับต่างๆนี้ โดยในส่วนของภาษาเขียนนั้น ระดับรูปที่เล็กที่สุดคือตัวอักษรแต่ละตัว จากตัวอักษรแต่ละตัวเราอาจจะพัฒนาระบบโดยให้พิจารณาเรื่องของอักขรวิธีหรือการผสมอักษรว่ามีกฎเกณฑ์หรือข้อห้ามอะไรบ้าง เช่น สระหน้าจะไม่ปรากฏติดกัน (เ-ไ) สระอือต้องมีตัวอักษรสะกดเสมอ (มื) รูปวรรณยุกต์ตรีจะไม่ปรากฏในพยางค์ที่มีพยัญชนะต้นเป็นอักษรสูง (ข๊า) เป็นต้น ในส่วนวจีวิภาค จะต้องหาสายอักขระที่มารวมกันเป็นหน่วยคำในภาษา ซึ่งเริ่มด้วยการหารายการหน่วยคำทั้งหมดที่เป็นไปได้ในภาษานั้น จากนั้นก็ต้องมีกฎเกณฑ์ที่ระบุการรวมหน่วยคำต่างๆเข้าด้วยกัน เพื่อใช้พิจารณาว่าส่วนใดเป็นราก ส่วนใดเป็นวิภัติปัจจัย เช่น หน่วยคำ "การ" เมื่อรวมกับหน่วยคำกริยาเช่น "เดิน" ก็จะทำให้เกิดหน่วยใหม่ที่เป็นคำนาม "การเดิน" เมื่อพัฒนาระบบในระดับวจีวิภาคได้แล้ว จึงพัฒนาระบบในระดับวากยสัมพันธ์โดยกำหนดกฎเกณฑ์ต่างๆทางวากยสัมพันธ์ให้กับระบบ เช่น การให้กฎต่างๆทางโครงสร้างของภาษานั้น ว่าโครงสร้างของประโยคประกอบด้วยส่วนใดบ้าง โครงสร้างของนามวลีประกอบด้วยส่วนใดบ้าง โครงสร้างกริยาวลีประกอบด้วยส่วนใดบ้าง เป็นต้น ซึ่งการที่คอมพิวเตอร์จะสามารถวิเคราะห์โครงสร้างประโยคเหล่านี้ได้ คอมพิวเตอร์จะต้องรู้ว่าคำในแต่ละคำนั้นจัดอยู่ในหมวดคำประเภทใด เช่น เป็นคำนาม คำ

กริยา คำวิเศษณ์ คำบุรพบท เป็นต้น วิธีหนึ่งที่สามารถทำได้คือการจัดเก็บคลังศัพท์ไว้ในคอมพิวเตอร์ โดยในคลังศัพท์นี้จะให้ข้อมูลว่าในภาษานั้นมีคำอะไรบ้าง คำนั้นจัดอยู่ในหมวดใด และอาจให้ข้อมูลอื่นๆเท่าที่จำเป็นต้องใช้สำหรับการประมวลผลภาษาธรรมชาติ เช่น ให้ข้อมูลว่าคำนั้นแทนมโนทัศน์อะไร ใช้ร่วมกับมโนทัศน์อะไรได้หรือไม่ได้ เช่น คำที่แทนมโนทัศน์ว่า +EAT จะเกิดกับประธานในกลุ่มคำที่มีมโนทัศน์เป็นสิ่งมีชีวิตประเภทสัตว์ +ANIMATE เป็นต้น การวิเคราะห์ในระดับอรรถศาสตร์จะได้ประโยชน์จากการใช้ข้อมูลทางอรรถศาสตร์เหล่านี้ นอกจากนี้ อาจต้องมีการจัดสร้างคลังความรู้ทั่วๆไปเก็บไว้ในคอมพิวเตอร์ด้วย เพื่อให้คอมพิวเตอร์สามารถตีความประโยคที่วิเคราะห์ได้ เช่น อาจกำหนดว่ามโนทัศน์ +HUMAN เป็นมโนทัศน์อยู่ภายใต้ +ANIMATE เป็นต้น ซึ่งจะทำให้คอมพิวเตอร์เข้าใจได้ว่าการที่คำกริยา +EAT เกิดร่วมกับประธานที่เป็น +ANIMATE ได้นั้น หมายความว่า คำกริยา +EAT สามารถเกิดร่วมกับประธานที่เป็น +HUMAN ได้ด้วย

นอกจากนี้ การที่จะทำให้คอมพิวเตอร์เข้าใจข้อความต่อเนื่องได้นั้น คอมพิวเตอร์จะต้องสามารถจดจำข้อมูลประโยคก่อนๆได้ด้วย ข้อมูลที่ได้จากประโยคใหม่ๆจะต้องนำไปเสริมความเข้าใจเดิมที่เกิดขึ้น จึงต้องออกแบบระบบเพื่อแทนแบบจำลองปริเฉจ และต้องมีกลไกที่ช่วยในการตีความคำสรรพนาม เช่น เขา เธอ ท่าน หรือคำนามชี้เฉพาะ เช่น ผู้ชายคนนั้น หนังสือเล่มนี้ ว่าอ้างถึงสิ่งใดที่เคยกล่าวมาถึงก่อนในบริบทนั้นหรือไม่ เพื่อจะได้ทำความเข้าใจกับความต่อเนื่องของข้อความที่ได้รับ

<u>พัฒนาการของการประมวลผลภาษาธรรมชาติ</u>

เมื่อได้เห็นภาพโดยคร่าวๆของระบบการประมวลผลภาษาธรรมชาติที่ควรจะเป็นแล้ว ในส่วนนี้เราจะมาพิจารณาว่าพัฒนาการของงานทางด้านนี้ที่เกิดขึ้นจริงเป็นอย่างไรบ้าง ได้พัฒนาโดยสอดคล้องกับแนวคิดในทางภาษาศาสตร์หรือไม่ เมื่อพิจารณาดู จะเห็นว่าพัฒนาการในงานด้านนี้เกิดจากนักคอมพิวเตอร์เป็นหลัก พัฒนาการที่มีมาจึงออกมาในรูปของการพัฒนาออกแบบแบบจำลองต่างๆ เช่น finite state automata, finite state transducer, hidden Markov model การพัฒนาเทคนิคการแจงส่วนประโยค (parse) เช่น การแจงส่วนแบบบนลงล่าง (top-down parsing) การแจงส่วนแบบล่างขึ้นบน (bottom-up parsing) เป็นต้น การพัฒนาอัลกอริทึมที่ใช้ เช่น dynamic programming, viterbi algorithm, และการพัฒนาแบบจำลองภาษาแบบต่างๆ เช่น การใช้ context-free grammar การใช้ feature structure เป็นต้น Jurafsky and Martin (2000) ได้สรุปภาพรวมของพัฒนาการทางด้านการประมวลผลภาษาธรรมชาติ โดยแยกออกเป็นยุคต่างๆ ดังนี้

1. ยุคแห่งการวางรากฐาน (1940's and 1950's)

งานทางด้านการประมวลผลภาษาธรรมชาติเริ่มต้นตั้งแต่ยุคหลังสงครามโลกครั้งที่สอง ในยุคนี้ มีงานต่างๆที่เป็นรากฐานของกระบวนทัศน์ (paradigm) ของการประมวลผลภาษาธรรมชาติที่ใช้กันอยู่ในปัจจุบัน คือ กระบวนทัศน์แบบสัญญะและกระบวนทัศน์แบบสถิติ งานที่เป็นรากฐานของการประมวลผลภาษาธรรมชาติ ได้แก่ แนวคิดเรื่อง automata และแนวคิดเกี่ยวกับแบบจำลองที่ใช้ความน่าจะเป็น (probabilistic model)

แนวคิดเรื่อง automata เกิดขึ้นในช่วงปี 1950 โดยได้รับอิทธิพลแนวคิดมาจากงานสามด้าน คือ Turing machine, Information Theory ของ Shannon และ neuron network ของ McCulloch-Pitts โดยที่ Turing machine เป็นแบบจำลองของกลไกสมมติที่ประกอบด้วยสภาพ (state) จำนวนจำกัดและมีเทปสำหรับเป็นอินพุท เอาท์พุท ในแต่ละสภาพ turing machine จะอ่านตัวอักษรจากอินพุทเทป 1 ตัว เพื่อประมวลผล จากนั้นจะเขียนตัว

อักษรที่ได้จากการประมวลผลลงในเอาท์พุทเทป จากนั้นอาจมีการเปลี่ยนไปสู่สภาพใหม่ พร้อมทั้งสามารถเลื่อนหัวอ่านไปทางซ้ายหรือขวาก็ได้ ส่วน Information Theory ของ Shannon นั้นนำแนวคิดของ finite state machine ซึ่งคล้ายคลึงกับ finite state automata มาใช้ในการศึกษาการส่งข้อมูลผ่านสื่อ พร้อมทั้งผนวกเอาแนวคิดเรื่องสถิติของ Markov มาใช้ด้วย โดยกำหนดให้แบบจำลองมีค่าความน่าจะเป็นในการเปลี่ยนสภาพหนึ่งไปสู่อีกสภาพหนึ่ง และ Shannon ได้ปรับปรุงแบบจำลองนี้ให้สามารถสร้างเอาท์พุทได้ในระหว่างการเปลี่ยนสภาพ แนวคิดที่สาม คือ neuron network ของ McCulloch-Pitts ซึ่งเป็นแบบจำลองของนิวรอน โดยนิวรอนแต่ละตัวเป็นเสมือนอุปกรณ์ไบนารี (binary device) ที่เชื่อมต่อกันคล้ายกับเครือข่ายของเซลสมอง แต่ละนิวรอนสามารถรับอินพุทจากนิวรอนที่อยู่ติดกัน และสามารถส่งผ่านเอาท์พุทไปยังนิวรอนตัวอื่นๆได้ถ้าค่าที่ประมวลผลได้นั้นเกินขีดขวางกั้น (threshold) แบบจำลองนี้เป็นต้นแบบสำหรับการศึกษาในเรื่อง neuron network และ connectionism ในเวลาต่อมา

ในปี 1956 Chomsky พยายามใช้แนวคิดจากงานของ Shannon เพื่อสร้างแบบจำลองทางไวยากรณ์ (grammar model) โดยที่ภาษาที่ได้จากแบบจำลองนี้เรียกว่า finite-state language ซึ่งต่อมาได้นำไปสู่การพัฒนาทฤษฎี formal language ขึ้นมาซึ่งรวมถึงแบบจำลองไวยากรณ์ที่เราเรียกว่า context-free grammar ด้วย และก็เป็นเหตุบังเอิญที่ context-free grammar นี้มีรูปแบบเดียวกับสิ่งซึ่ง Backus (1959) และ Naur et al (1960) สร้างขึ้นมาเพื่อใช้อธิบายกฎเกณฑ์ของภาษาคอมพิวเตอร์ ALGOL

ส่วนแบบจำลองที่ใช้แนวคิดของสถิติความน่าจะเป็นนั้นเริ่มต้นมาจากงานของ Shannon ที่นำเอาสถิติและแนวคิดเรื่องเอ็นโทรปีในเทอร์โมไดนามิคมาใช้ในแบบจำลองการสื่อสารของเขา งานในระยะแรกที่ออกมาจากการใช้วิธีการทางสถิตินี้คือระบบรู้จำตัวเลขของบริษัทเบลล์

2. ยุคแบ่งเป็นสองค่าย (The Two Camps) 1957-1971

ในช่วงปลายของทศวรรษ 1950 และต้นทศวรรษ 1960 มีการแบ่งแยกแนวคิดเรื่องการประมวลผลวัจนะและการประมวลผลภาษา (speech and language processing) ออกจากกันเป็น 2 กระบวนทัศน์อย่างชัดเจน คือ การประมวลผลวัจนะจะใช้วิธีการทางสถิติ (stochastic) ส่วนการประมวลผลภาษาจะใช้แนวคิดแบบสัญญะ (symbolic)

ในกระบวนทัศน์แบบสัญญะมีงานที่ออกมาในสองกลุ่มใหญ่ๆ คือ กลุ่มแรกเป็นงานของนักภาษาศาสตร์ในสายของ formal linguistics เช่นกลุ่มของ Chomsky และรวมไปถึงงานทางด้านออกแบบตัวแปลภาษา (compiler) ของภาษาคอมพิวเตอร์ และการพัฒนาโปรแกรมแจงส่วนประโยค ส่วนกลุ่มที่สองเป็นงานทางด้านปัญญาประดิษฐ์ของนักคอมพิวเตอร์ ซึ่งให้ความสำคัญกับเรื่องของการหาเหตุผลและตรรกวิทยา โปรแกรม Eliza ที่กล่าวถึงในตอนต้นก็เกิดขึ้นในยุคนี้

ส่วนกระบวนทัศน์แบบสถิติมีรากฐานพัฒนามาจากภาควิชาวิศวกรรมไฟฟ้าและภาควิชาสถิติ ในปลายทศวรรษ 1950 มีการประยุกต์ใช้ Bayesian method กับปัญหาเรื่องการรู้จำตัวอักษร (optical character recognition) และมีการใช้ Bayesian method สำหรับการรู้จำเอกสาร (text recognition) เพื่อดูว่าใครเป็นคนแต่ง ต่อมามีการขยายงานของ Markov และ Shannon ในช่วงปลายทศวรรษที่ 1960 ออกมาเป็น hidden Markov model เพื่อใช้กับปัญหาการรู้จำเสียง (speech recognition)

กล่าวได้ว่าอัลกอริทึมพื้นฐานที่สำคัญในกระบวนทัศน์แบบสัญญะและแบบสถิติที่ใช้กันในการประมวลผลภาษาธรรมชาติในปัจจุบันนั้นมีปรากฏมาตั้งแต่ช่วงต้นทศวรรษ 1970 นี้

3. ยุคของการประมวลผลภาษาธรรมชาติ : 1972-1983

ในยุคนี้มีการพัฒนาอัลกอริทึมหลักต่าง ๆ ที่เป็นที่รู้จักกัน เช่น ในปี1972 Winograd สร้างระบบที่เรียกว่า SHRDLU ซึ่งเป็นการสร้างแบบจำลองหุ่นยนต์ในโลกที่มีกล่องของเล่นรูปร่างต่างๆจำนวนหนึ่ง โปรแกรม SHRDLU สามารถเข้าใจคำสั่งภาษาอังกฤษเช่น “Move the red block on the top of the smaller green one” ว่าเป็นคำสั่งให้หยิบของสิ่งใด งานของ Winograd ทำให้นักวิจัยในสาขารู้สึกว่าความรู้ความเข้าใจในเรื่องของการแจงส่วนประโยคนั้นใช้ได้ดีพอแล้ว จึงมีผู้เริ่มสนใจศึกษาในเรื่องของความหมาย ซึ่งในช่วงทศวรรษที่ 1970 ก็มีงานจำนวนมากที่เกี่ยวข้องกับความหมายและการประมวลผลข้อความต่อเนื่อง เช่น งานของ Roger Schank แห่งมหาวิทยาลัยเยลที่พยายามสร้างแบบจำลองการเข้าใจภาษาโดยเน้นที่เรื่องของการสร้างรูปแทนความหมายและการสร้างฐานความรู้เกี่ยวกับโลก (world knowledge) อย่าง script, plans and goals ส่วนงานที่สร้างแบบจำลองทางความหมายโดยอาศัย predicate logic คืองานของ Woods ในปี 1978 ที่สร้างระบบถามตอบ LUNAR (question-answering system)

และในช่วงนี้ ก็มีงานที่เกี่ยวกับแบบจำลองของการประมวลผลปริเฉจ (discourse processing) เช่น Grosz พัฒนาแนวคิดเรื่อง discourse focus เพื่อนำมาใช้แก้ปัญหาการอ้างถึงของคำสรรพนาม คำนามชี้เฉพาะ นอกจากนี้ ในยุคนี้ ก็เกิดงานที่ใช้แนวคิดเรื่อง unification operation เช่น การพัฒนาทฤษฎี definite clause grammar และทฤษฎี Lexical Functional Grammar

หลังจากปี 1983 มาจนถึงช่วงต้นทศวรรษ 1990 งานด้านการประมวลผลภาษาไม่มีการเปลี่ยนแปลงหรือสร้างสรรค์แนวคิดใหม่ๆ ออกมามากนัก และศาสตร์ด้านการประมวลผลภาษาดูจะไม่ได้รับความนิยมหรือความสนใจมากนัก เนื่องมาจากการที่ไม่สามารถสร้างระบบที่สามารถทำงานได้จริงจังอย่างที่คาดหวัง จนบางคนเริ่มสงสัยว่าความฝันอันนี้จะเป็นจริงขึ้นมาได้หรือไม่

4. ยุคปัจจุบัน รวมกระบวนทัศน์เป็นหนึ่งเดียว (1994-1999)

ในยุคปัจจุบันนี้ได้ปรากฏชัดว่า กระบวนทัศน์แบบสถิติซึ่งใช้กันเป็นหลักในการประมวลผลวัจนะเริ่มได้รับการยอมรับนำไปใช้ในการประมวลผลภาษาด้วย อัลกอลิทึมต่างๆที่ใช้เริ่มหันมานำแนวคิดด้านสถิติหรือเรื่องความน่าจะเป็นเข้ามาประยุกต์ใช้ นอกจากนี้ การเติบโตอย่างรวดเร็วของอินเตอร์เน็ตทำให้มีความต้องการระบบที่สามารถทำงานค้นคืนข้อมูล (information retrieval) และคัดใจความเอกสาร (information extraction) ได้จากเอกสารที่มีอยู่มากมายในอินเตอร์เน็ต จึงจำเป็นต้องอาศัยระบบที่สามารถประมวลผลภาษาจริงๆตามที่ปรากฏในเอกสารและมีความหลากหลายได้ วิธีการทางสถิติช่วยให้สามารถพัฒนาระบบที่ตอบสนองความต้องการนี้ได้ดีกว่าการใช้วิธีการกำหนดกฎตามกระบวนทัศน์แบบสัญญะ

หากพิจารณาพัฒนาการของระบบการประมวลผลภาษาธรรมชาติที่กล่าวมา จะเห็นว่าการพัฒนางานด้านนี้ส่วนมากไม่ได้สัมพันธ์โดยตรงกับพัฒนาการของภาษาศาสตร์นัก ทั้งนี้อาจเป็นเพราะจุดมุ่งหมายที่แตกต่างของศาสตร์ทั้งสอง คือในขณะที่นักภาษาศาสตร์สนใจหาสิ่งที่เป็นความรู้เกี่ยวกับภาษา (knowledge of language) เพื่อตอบคำถามว่าภาษาคืออะไร ทำไมมนุษย์เราจึงมีความสามารถทางภาษา สามารถเรียนรู้และเข้าใจภาษาได้

ลักษณะร่วมของมนุษย์ที่มีมาแต่กำเนิดที่ทำให้สามารถเรียนรู้ภาษาได้หรือสมรรถนะทางภาษา (linguistic competence) คืออะไร ดังนั้น การศึกษาของนักภาษาศาสตร์จึงพุ่งไปในประเด็นปลีกย่อยต่างๆ เช่น ทำไมประโยคแบบนี้จึงไม่ถูกไวยากรณ์ ทำไมประโยคแบบนี้จึงถูกไวยากรณ์ หากสลับที่บางคำหรือเพิ่มเติมบางส่วน ประโยคเหล่านี้จะยังคงถูกต้องตามไวยากรณ์หรือไม่ และทั้งหมดนี้ควรจะอธิบายด้วยแบบจำลองทางไวยากรณ์แบบใดจึงจะเหมาะสมที่สุด ซึ่งแบบจำลองนั้นก็ควรจะต้องมีรากฐานมาจากหลักการร่วมกันที่เป็นสากลหรือสามารถใช้ได้กับทุกๆ ภาษาในโลกนี้ จนถึงปัจจุบัน นักภาษาศาสตร์ยังคงค้นคว้าเพื่อหาคำตอบเหล่านี้อยู่ ส่วนนักคอมพิวเตอร์นั้น เนื่องจากความสนใจมุ่งอยู่ที่การสร้างระบบหรือโปรแกรมที่สามารถทำงานด้านภาษาได้ หากจะต้องการฐานความรู้ทางภาษาก็เพื่อตอบสนองวัตถุประสงค์ดังกล่าว และเนื่องจากความรู้ที่จะนำมาประยุกต์ใช้ได้นั้น จำเป็นต้องใช้ความรู้ที่มีแบบแผนชัดเจน (formal) งานของนักภาษาศาสตร์ที่สามารถนำมาประยุกต์ใช้ได้จึงเป็นทฤษฎีในกลุ่มแบบแผนนิยม (formal grammar) ทฤษฎีภาษาศาสตร์ในบางกลุ่มจึงไม่สามารถนำมาใช้ได้โดยตรง เช่น ทฤษฎีไวยากรณ์หน้าที่นิยม (functional grammar) ตามแนวทางของ Givon ทฤษฎีที่มีการนำมาใช้ในการพัฒนาระบบการประมวลผลภาษาธรรมชาติ เช่น Head-driven Phrase Structure Grammar, Generalized Phrase Structure Grammar, Lexical Functional Grammar เป็นต้น แต่การที่จะทำให้ระบบทำงานได้กับข้อมูลภาษาจริงๆแล้ว ไวยากรณ์ที่ใช้นั้นจะต้องสมบูรณ์พอที่จะครอบคลุมข้อมูลเอกสารนั้นๆ ซึ่งเป็นสิ่งที่นักคอมพิวเตอร์ผู้พัฒนาระบบต้องหากฎเหล่านี้ออกมาใช้งานเองด้วย เพราะการหารายการกฎให้ครอบคลุมทุกๆประโยคนั้นไม่ใช่สิ่งที่นักภาษาศาสตร์ส่วนใหญ่สนใจทำ

นอกจากนี้ งานในกระบวนทัศน์แบบสถิติของการประมวลผลภาษาธรรมชาตินั้น เป็นสิ่งที่ไม่ได้อยู่ในความสนใจของนักภาษาศาสตร์ นักภาษาศาสตร์โดยทั่วไปจะมองว่าเป็นเพียงวิธีคิดแก้ปัญหาแบบง่ายๆของนักคอมพิวเตอร์ที่ไม่มีความรู้ทางภาษาศาสตร์ดีพอ แต่หากพิจารณาจากแนวคิดพื้นฐานที่มีอยู่ในทางภาษาศาสตร์แล้ว พัฒนาการของแนวคิดพื้นฐานในทางภาษาศาสตร์นั้น เราจะพบลักษณะที่สอดคล้องกับพัฒนาการของแนวคิดพื้นฐานในงานประมวลผลภาษาธรรมชาติ คือเป็นการศึกษาในสองแนวทางใหญ่ๆ คือ แนวทางแบบเหตุผลนิยม (rationalism) กับแบบปฏิบัตินิยม (empiricism) ในช่วงตั้งแต่ปี 1960 มาจนถึง 1985 งานทางด้านภาษาศาสตร์ จิตวิทยา ปัญญาประดิษฐ์รวมไปถึงการประมวลผลภาษาธรรมชาติเป็นไปตามแนวทางแบบเหตุผลนิยมเป็นหลัก ซึ่งเชื่อว่าความรู้ของมนุษย์มีลักษณะที่เหมือนกฎเกณฑ์ตายตัวที่อยู่ในหัว ในสาขาภาษาศาสตร์ แนวคิดแบบเหตุผลนิยมเข้ามาพร้อมกับการยอมรับแนวคิดของ Chomsky ดังที่ได้กล่าวไว้ในตอนต้น Chomsky เชื่อว่าสมรรถนะทางภาษามีลักษณะที่เป็น innate โดยยกเหตุผลในเรื่องของอินพุทที่ไม่สมบูรณ์ คือบอกว่าเป็นไปได้ยากที่เด็กจะเรียนรู้ระบบของภาษาที่ซับซ้อนได้จากอินพุทที่ไม่สมบูรณ์นี้ได้ คนเราไม่ได้เรียนรู้ภาษาโดยการเลียนแบบพฤติกรรม แต่มีความสามารถในการสร้างประโยคแบบใหม่ๆโดยที่ไม่เคยได้ยินมาก่อน อินพุทที่ไม่สมบูรณ์เหล่านั้นเป็นเพียงตัวกระตุ้นให้เกิดการสร้างระบบไวยากรณ์ของภาษา ในขณะที่พวกปฏิบัตินิยมไม่เชื่อว่าคนเรามีกลไกทางภาษาติดตัวมาแต่กำเนิด เพียงแต่ว่ามีกลไกพื้นฐานสำหรับการรับรู้ (cognition) ทั่วๆไป เช่น การโยงความสัมพันธ์ (association) การรู้จำรูปแบบ (pattern recognition) การสรุปรูปแบบ (generalization) ซึ่งเป็นสิ่งที่เด็กจะสามารถใช้ในการเรียนรู้และสร้างระบบภาษาขึ้นมาจากข้อมูลภาษาที่ได้รับ ความจริงแนวคิดแบบปฏิบัตินิยมนี้มีมาตั้งแต่ช่วงปี 1920 ถึง 1960 คือก่อนหน้ายุคของ Chomsky และเพิ่งเริ่มกลับมาได้รับความสนใจจากนักภาษาศาสตร์บาง

ส่วนในช่วงปี 1990 นี้ ส่วนงานทางด้านปัญญาประดิษฐ์ของคอมพิวเตอร์ซึ่งพยายามสร้างระบบที่ดูเสมือนว่ามีความฉลาดเหมือนกับมนุษย์ ซึ่งเริ่มตั้งแต่ช่วงปี 1970-1989 นั้นก็ใช้แนวคิดแบบเหตุผลนิยมในการพัฒนาระบบ แต่ระบบต่างๆที่พัฒนามานั้นปัจจุบันถูกวิจารณ์ว่าเป็นเสมือนของเล่น เพราะสามารถใช้ได้เฉพาะกับปัญหาเล็กๆในขอบเขตที่จำกัด ต่อมาเมื่อมีความต้องการที่จะทำให้ระบบการประมวลผลภาษาสามารถทำงานกับข้อมูลเอกสารจริงๆได้ ระบบที่ใช้สถิติช่วยจึงเป็นที่นิยมยอมรับอย่างแพร่หลายมากกว่าเพราะสามารถทำงานได้ดีและสะดวกกว่าระบบแบบเดิมที่ใช้การกำหนดกฎเกณฑ์ต่างๆด้วยคน การนำเอาสถิติเข้ามาใช้นี้โดยพื้นฐานเป็นการยอมรับแนวคิดแบบปฏิบัตินิยมคือมองว่า คนเราสามารถเรียนรู้ภาษาที่มีความสลับซับซ้อนมากมายได้โดยมีเพียงแบบจำลองพื้นฐานทางภาษา (general language model) แล้วเรียนรู้ค่าพารามิเตอร์ต่างๆจากการใช้ค่าทางสถิติและใช้วิธีการเรียนรู้แบบที่ใช้ใน machine learning ซึ่งในงานประมวลผลภาษานี้ ก็มักจะสมมติใช้เอกสารจากคลังข้อมูลภาษาว่าเป็นเสมือนอินพุทภาษาที่คนเราได้ยินได้ฟัง ความจริงแนวคิดในลักษณะนี้ไม่ได้แตกต่างไปจากแนวคิดที่นักภาษาศาสตร์ตามแนวไวยากรณ์โครงสร้างได้เสนอไว้แล้วตั้งแต่ปี 1951 ในเรื่องของ discovery procedure คือการพยายามที่จะค้นพบระบบของภาษาโดยอัตโนมัติจากการศึกษาเปรียบเทียบข้อมูลภาษาที่รวบรวมได้

คำถามที่น่าสนใจ คือการนำสถิติเข้ามาใช้กับการภาษานั้นเป็นเพียงการแก้ปัญหาแบบง่ายๆของนักคอมพิวเตอร์อย่างเดียวจริงหรือไม่ สถิติมีส่วนเกี่ยวข้องสัมพันธ์กับภาษาศาสตร์เพียงใด แม้ระบบการประมวลผลภาษาที่ใช้วิธีการแบบสถิติจะทำงานได้ผลดีมากเพียงใด นักภาษาศาสตร์ก็จะแย้งว่า วิธีการแบบนี้ไม่ได้อธิบายกลไกการเข้าใจภาษาของมนุษย์ อย่างไรก็ตาม ในเรื่องนี้ Abney (1996) มองว่า เรื่องทางสถิตินั้นมีส่วนเกี่ยวข้องและช่วยให้เราเข้าใจถึงกลไกของภาษาได้ดีขึ้น โดยเขาได้ชี้ให้พิจารณาในสามประเด็นในเรื่องของภาษา คือ การเรียนรู้ภาษา (language acquisition) การเปลี่ยนแปลงของภาษา (language change) และการแปรของภาษา (language variation) ว่ามีความเกี่ยวข้องกับเรื่องสถิติ

ในเรื่องการเรียนรู้ภาษานั้น ถ้าเด็กเรียนรู้ไวยากรณ์ในแบบที่เป็นกฎเกณฑ์แบบพีชคณิต (algebraic) คือมีลักษณะเป็นกฎตายตัวว่าอะไรใช่หรือไม่ใช่ ถูกหรือผิด เราคงจะเห็นการเปลี่ยนแปลงของการใช้ภาษาของเด็กอย่างเฉียบพลัน กฎบางอย่างจะเกิดขึ้นและกฎบางอย่างจะหายไปในวันนี้วันพรุ่งนี้ แต่ในความจริง การเปลี่ยนแปลงทางภาษาของเด็กเป็นการเปลี่ยนแปลงในเชิงความถี่สัมพัทธ์ทางโครงสร้าง ซึ่งน่าจะเป็นว่า เด็กจะมีกฎของหลายๆโครงสร้างที่แข่งขันกันอยู่ภายใน สิ่งที่กำหนดว่าโครงสร้างไหนถูกต้องมากกว่ากันสามารถมองในลักษณะของค่าความน่าจะเป็นซึ่งสามารถเปลี่ยนแปลงค่าไปได้เรื่อยๆในระหว่างการเรียนรู้ภาษา จนกระทั่งโครงสร้างใดที่มีค่าความน่าจะเป็นเป็นศูนย์ เด็กก็จะเลิกใช้โครงสร้างนั้นไป ลักษณะนี้เป็นการเพิ่มคุณสมบัติของสถิติเข้าไปในส่วนของไวยากรณ์ทำให้ไวยากรณ์มีลักษณะที่เป็นแบบสถิติ (stochastic) แทนที่จะเป็นแบบพีชคณิต

ในเรื่องการเปลี่ยนแปลงของภาษานั้นก็เป็นเช่นเดียวกัน ถ้ากฎในภาษามีลักษณะเป็นแบบพีชคณิตเราก็คงจะเห็นการเปลี่ยนแปลงทางภาษาอย่างทันทีทันใด แต่ความจริงไม่ได้เป็นเช่นนั้น การเปลี่ยนแปลงของภาษาเป็นเรื่องที่ใช้เวลาเป็นสิบเป็นร้อยปี ไม่ใช่ว่า อยู่ๆเราเดินไปที่ร้านเหล้าแห่งหนึ่ง แล้วสั่ง ale (ออกเสียง /al/) ซึ่งเป็นเบียร์ชนิดหนึ่งกลับได้ eel (ออกเสียง /el/) หรือปลาไหลมาแทนเพราะว่าเมื่อวานนี้ตอนที่เราออกนอกเมืองไปได้เกิดการเลื่อนสระจาก /e/ เป็น /a/ ไปแล้ว แต่ถ้าหากเรามองว่า ภาษาในกลุ่มสังคมหนึ่งๆเป็นภาพรวมเชิงสถิติของภาษาของปัจเจกในสังคมนั้นๆ เช่น ค่าความน่าจะเป็นของโครงสร้างหนึ่งในสังคมนั้นจะเท่ากับสัดส่วนของคนที่ใช้โครง

สร้างนั้นต่อคนทั้งหมด ในมุมมองแบบนี้ เราก็สามารถมองการเปลี่ยนแปลงของภาษาในลักษณะที่ค่อยๆเป็นการเปลี่ยนแปลงในลักษณะที่เลื่อนออกจากศูนย์กลางของคนในสังคมนั้นๆได้

แต่คำถามที่ต้องถามต่อไปก็คือ แล้วสถิติเกี่ยวข้องกับไวยากรณ์ของคนแต่ละคนที่อยู่ในสังคมที่พูดภาษาที่เหมือนกัน (monolingual speaker in homogeous speech community) หรือไม่ เพราะนี่คือบริบทที่ Chomsky อ้างถึงในการสร้างแบบจำลองของสมรรถนะทางภาษา (language competence) ถ้าพิจารณาอย่างผิวเผิน คำตอบก็คือไม่น่าจะเกี่ยวข้องกัน เรื่องทางสถิติไม่เกี่ยวข้องกับสิ่งที่นักภาษาศาสตร์ส่วนใหญ่กำลังศึกษาอยู่ ซึ่งเป็นเรื่องของการหากฎเกณฑ์ (principle) ที่อยู่ภายในของ language faculty ซึ่งเป็นแบบจำลองในอุดมคติ (idealization) ของคนที่อยู่ในสังคมที่พูดภาษาที่เหมือนกัน แต่ Abney ได้ตั้งข้อสงสัยว่าสิ่งที่นักภาษาศาสตร์สนใจนี้แคบเกินไปหรือเปล่าในสภาพการณ์ปัจจุบันนี้ แบบจำลองของไวยากรณ์ที่นักภาษาศาสตร์สร้างขึ้นสามารถอธิบายภาษาได้ในขอบเขตแคบๆเท่านั้น ข้อมูลอื่นๆที่ไม่สามารถอธิบายได้หรืออยู่นอกเหนือจากกฎเกณฑ์มักจะถูกมองว่าเป็นเรื่องของการใช้ภาษา (performance) ไม่ใช่เรื่องของ competence และออกตัวว่าอยู่นอกเหนือจากขอบเขตของการศึกษาทางภาษาศาสตร์ กล่าวคือ นักภาษาศาสตร์ไม่ได้ให้ความสำคัญกับข้อมูลจริงมากนัก แต่กลับสนใจอยู่กับข้อมูลสังเคราะห์ซึ่งไม่พบเห็นในชีวิตจริง และบางครั้งก็ยากที่ทุกคนจะตัดสินออกมาในลักษณะเดียวกันว่าเป็นประโยคที่ถูกไวยากรณ์หรือไม่ เช่น

วันนี้ ฉันเห็นนักเขียน$_i$ที่นิดบอกหน่อยว่าแดงกำลังอ่านหนังสือที่ e_i วิจารณ์

Whom$_i$ do you think that Bill said that John thought that Harry shaved e_i ?

*Who$_i$ did you say was bothered by our talking to e_i ?

และในความเป็นจริง มีข้อมูลมากมายที่ไวยากรณ์แบบที่ใช้กัน คือไวยากรณ์ในเชิงพีชคณิตไม่สามารถอธิบายได้ จริงๆแล้วคุณสมบัติแบบพีชคณิตอาจจะเป็นเพียงคุณสมบัติหนึ่งในหลายๆคุณสมบัติของภาษา เป็นไปได้ไหมว่าเราอาจจะกำลังมองข้ามแง่มุมบางอย่างของภาษาไป Abney กล่าวสรุปในเรื่องนี้ว่านักภาษาศาสตร์ได้หลงลืมจุดมุ่งหมายเดิม และการเดินตามแนวคิดของ Chomsky เป็นการปิดหนทางการศึกษาภาษาอย่างแท้จริง (linguistics has lost sight of its original goal, and turned Chomsky's expedient into an end in itself)

<u>ทำไมการประมวลผลภาษาจึงเป็นเรื่องยาก</u>

การประยุกต์ใช้ไวยากรณ์แบบพีชคณิตในระบบการประมวลผลภาษาธรรมชาติทำให้เราเริ่มเห็นว่าไวยากรณ์แบบพีชคณิตที่ใช้กันนั้นอาจไม่สอดคล้องกับการรับรู้ของมนุษย์ และเป็นสาเหตุที่ทำให้การประมวลผลภาษาเป็นเรื่องยาก กล่าวคือในประโยคหนึ่งๆที่วิเคราะห์นั้นจะมีโครงสร้างที่เป็นไปได้มากมาย ก่อให้เกิดปัญหาในเรื่องของความกำกวมซึ่งระบบจะต้องตัดสินว่าโครงสร้างใดประโยคแบบใดควรจะเป็นโครงสร้างที่ต้องการ หากลองพิจารณาตัวอย่างประโยคอย่าง Companies are training workers ซึ่งคนเรารับรู้และเข้าใจ (perceive) แบบเดียว คือ ในแบบที่ are training เป็นกลุ่มของกริยา แต่ระบบประมวลผลภาษาสามารถมองเห็นเพิ่มอีกสองแบบ คือ แบบที่ are เป็นกริยาหลักและส่วนที่เหลือเป็น gerund เหมือนอย่างในประโยค Our problem is training workers หรือแบบที่ are เป็นกริยาหลักและส่วน training เป็นส่วนขยายของ workers เหมือนอย่างในประโยค Those are training wheels หรือในตัวอย่างประโยค List the sales of the products produced in 1973 with the products produced in 1972 (Manning and Schutze 1999: 9) ซึ่งเป็นประโยคที่เราคาดหวังจะใช้สื่อกับคอมพิวเตอร์ ใน

ประโยคนี้ มีผู้รายงานว่า ระบบประมวลผลภาษาธรรมชาติของเขาสามารถแจงส่วนได้ถึง 455 แบบ การพยายามที่จะแก้ปัญหานี้หรือการพยายามลดความกำกวมที่เกิดขึ้น โดยใช้วิธีเพิ่มกฎทางไวยากรณ์ก็จะส่งผลกระทบต่อประโยคอื่นๆ ทำให้ความสามารถในการแจงส่วนประโยคอื่นๆลดลง ในขณะที่ถ้าพยายามลดกฎทางไวยากรณ์ให้ครอบคลุมประโยคมากขึ้นก็จะส่งผลให้มีความกำกวมเกิดขึ้นมากตามไปด้วย ปัญหาเหล่านี้ เราจะพบมากเวลาที่แจงส่วนประโยคต่างๆ ประโยคที่ดูเหมือนเป็นประโยคง่ายๆและไม่มีความกำกวมทางโครงสร้าง แต่ระบบประมวลผลภาษาธรรมชาติจะพบว่ามีโครงสร้างที่เป็นไปได้มากกว่าหนึ่งโครงสร้าง คำถามคือ จะทำอย่างไรที่จะทำให้ระบบประมวลผลภาษาธรรมชาตินั้นสามารถเลือกโครงสร้างให้สอดคล้องกับแบบที่คนเรารับรู้ได้ วิธีการหนึ่งก็คือ การนำเรื่องของความน่าจะเป็นเข้ามาประยุกต์ใช้ โดยกำหนดให้โครงสร้างแต่ละโครงสร้างมีค่าความน่าจะเป็นที่ไม่เท่ากัน โครงสร้างที่มีค่าความน่าจะเป็นสูงเกินค่าระดับหนึ่ง ก็ให้ถือว่าเป็นโครงสร้างที่คนเราสามารถรับรู้ได้ ปัญหาเรื่องความกำกวมที่เกิดขึ้นนี้เป็นอุปสรรคอย่างมากโดยเฉพาะสำหรับระบบประมวลผลภาษาธรรมชาติที่ใช้กระบวนทัศน์แบบสัญญะ ในขณะที่ระบบที่ใช้กระบวนทัศน์แบบสถิติจะจัดการกับปัญหานี้ได้ดีกว่า เพราะการใช้สถิติเรื่องความน่าจะเป็นจะช่วยในการเลือกว่าโครงสร้างแบบใดที่น่าจะถูกต้องมากที่สุด

<u>ไวยากรณ์แบบอิงสถิติ</u>

ปัญหาในเรื่องความกำกวมที่เกิดขึ้นจากการประยุกต์ใช้กฎไวยากรณ์แบบพีชคณิตที่นักภาษาศาสตร์ใช้กันนั้นชี้ให้เห็นว่าไวยากรณ์ในแบบที่ศึกษากันอยู่นั้นอาจจะไม่สอดคล้องกับการรับรู้ (perception) ของมนุษย์ คำถามหนึ่งที่เราสามารถถามตัวเองได้คือ “ประโยคหรือวลีนี้ฟังดูเป็นธรรมชาติไหม” (sound natural) ซึ่งการฟังดูเป็นธรรมชาติเป็นเรื่องที่เป็นระดับชั้น (degree) คือจะตอบว่าฟังดูเป็นธรรมชาติมากหรือน้อยไม่ใช่เรื่องที่จะตอบว่าใช่หรือไม่ใช่ ซึ่งลักษณะคำตอบเช่นนี้เราไม่สามารถอธิบายได้ด้วยกฎที่มีลักษณะตายตัวแบบพีชคณิตได้ แต่สามารถอธิบายได้ด้วยการใช้ไวยากรณ์ที่มีการถ่วงน้ำหนัก (weighted grammar) หรือไวยากรณ์เชิงสถิติคือ ไวยากรณ์ที่ใช้ค่าความน่าจะเป็นประกอบด้วย ซึ่งแบบจำลองของไวยากรณ์แบบนี้จะสอดคล้องกับวิจารณญาณของมนุษย์ (human judgment) มากกว่า กล่าวคือ ประโยคที่สร้างขึ้นจากไวยากรณ์นั้นนอกจากจะบอกได้ถึงเรื่องว่าถูกไวยากรณ์หรือไม่แล้วยังบอกถึงระดับชั้น (degree) ว่าประโยคเหล่านั้นฟังดูเป็นธรรมชาติเพียงใด และก็ควรจะต้องสอดคล้องกับการรับรู้ของมนุษย์

นอกจากนี้ แนวคิดทางสถิติยังช่วยอธิบายเรื่องของการเกิดร่วมกันของคำ (collocation) และข้อจำกัดของการเลือกเกิดร่วมกัน (selectional restriction) ได้ หรืออย่างน้อยก็บอกถึงระดับชั้นของความเป็นธรรมชาติ (degree of naturalness) ได้ เช่น บอกได้ว่า strong tea และ powerful car จะฟังดูเป็นธรรมชาติมากกว่า strong car กับ powerful tea เพราะจากค่าทางสถิติ จะพบว่ามีการปรากฏร่วมกันของ strong กับ tea มากกว่า strong กับ car ซึ่งบางคนอาจจะแย้งว่าเป็นการยากที่จะกำหนดว่า ความเป็นธรรมชาติ (naturalness) คืออะไร แต่ในความเป็นจริงความเป็นธรรมชาติเป็นสิ่งที่เราสามารถรับรู้ได้ ความเป็นธรรมชาติเป็นคนละเรื่องกับการมีความหมายที่เข้าใจได้ (meaningfulness) ตัวอย่างเช่น เรารู้สึกว่า differential structure ฟังดูเป็นธรรมชาติมากกว่า differential child ถึงแม้ว่าเราจะไม่รู้ว่า differential structure หมายถึงอะไรก็ตาม

นอกจากนี้ เรื่องทางสถิติยังสามารถนำมาใช้อธิบายปรากฏการณ์เรื่องอื่นๆในทางภาษาได้อีกด้วย ตัวอย่างเช่น (a) การปรับแก้ความผิดพลาด (error tolerance) เช่น เวลาที่ผู้ฟังได้ยินประโยค Thanks for all you help

ผู้ฟังมักเลือกที่จะตีความประโยคนี้ใหม่เป็น thanks for all your help มากกว่า ที่จะคิดถึงประโยคนี้ในโครงสร้างที่ถูกไวยากรณ์ซึ่งมีหมายความว่า thanks for all those who you help เรื่องนี้สามารถอธิบายได้ว่า เพราะความสิ้นเปลือง (cost) ของการแก้ข้อผิดพลาด (error correction) เพื่อให้ได้โครงสร้างที่รู้สึกดีกว่า (preferred) (ซึ่งมีค่าสถิติการใช้มากกว่า)นั้นน้อยกว่าความสิ้นเปลืองของการพยายามสร้าง (derive) โครงสร้างที่ถูกไวยากรณ์แต่มีการค่าสถิติการใช้น้อยกว่า (less preferred) (b) การเรียนรู้ในทันที (learning on the fly) เช่น เวลาที่ผู้ฟังได้ยินประโยคว่า a hectare is a hundred ares ผู้ฟังจะสรุปว่า ares เป็นคำนามได้โดยที่ไม่จำเป็นต้องรู้จักคำนี้มาก่อน การที่คนเราสามารถสร้างโครงสร้างประโยคที่คิดเอาไว้ได้ (pick intended structure) สามารถอธิบายได้ว่า คนเรามีกระบวนการที่จะเรียนรู้ โดยการกำหนดหมวดคำให้กับคำใหม่ เพิ่ม subcategorization frame ของคำกริยาที่เกี่ยวข้อง และคำนวณค่าความสิ้นเปลืองของการเรียนรู้นั้น ซึ่งการใช้ไวยากรณ์แบบมีน้ำหนักได้ หรือไวยากรณ์ที่มีการใช้ค่าความน่าจะเป็นก็เป็นวิธีหนึ่งที่สามารถนำมาอธิบายปรากฏการณ์นี้ได้

แนวคิดทางสถิติยังช่วยอธิบายในเรื่องของการเลือกโครงสร้างตามความนิยมใช้ (structural preference) ได้ เช่น ในประโยค the emergency crews hate most is domestic violence ในประโยคนี้นั้น โครงสร้างที่ถูกคือ the emergency เป็นนามวลีและ crews hate most เป็นส่วนขยายของนามวลี (the emergency (that) crews hate most) แต่เนื่องจากการเลือกตีความแบบที่ยาวที่สุดเป็นทางเลือกที่นิยมใช้มากกว่า ในการอ่านรอบแรก เราจึงอ่านนามวลีตัวแรกเป็น the emergency crews จนเมื่ออ่านต่อๆไปจนจบจึงพบว่าไม่ถูกต้อง ต้องย้อนกลับไปอ่านข้อความเดิมใหม่

<u>บทสรุป</u>

ไม่ว่าระบบประมวลผลภาษาธรรมชาติจะใช้กระบวนทัศน์แบบสถิติหรือแบบสัญญะก็ตาม ระบบการประมวลผลภาษาธรรมชาตินั้นก็ยังคงต้องอาศัยการพัฒนาโมดูลต่างๆตามแนวคิดทางภาษาศาสตร์ กล่าวคือ หากอินพุทที่รับเข้าเป็นข้อความต่อเนื่อง ระบบประมวลผลภาษาก็จะต้องมีโมดูลที่แยกแยะคำโดยทำการวิเคราะห์ระดับวจีวิภาค (morphological analysis) เพื่อหารูปคำและวิภัติปัจจัยที่มี จากนั้นมีการหาหมวดคำของคำแต่ละคำ (part-of-speech tagging) เพื่อเป็นข้อมูลพื้นฐานในการวิเคราะห์ทางวากยสัมพันธ์และทางความหมาย (syntactic and semantic analysis) จากนั้นจึงผนวกประโยคที่วิเคราะห์ได้เข้าเป็นส่วนหนึ่งของข้อความทั้งหมดที่ประมวลผลแล้วและเก็บไว้ในแบบจำลองปริเฉจ (discourse model) ซึ่งก็จะต้องมีประบวนการตรวจสอบการอ้างอิงว่าคำนามคำสรรพนามที่ใช้ในประโยคนั้นอ้างถึงสิ่งที่เคยอ้างถึงมาก่อนหรือไม่ (reference resolution) เมื่อระบบต้องการโต้ตอบกลับเป็นภาษาธรรมชาติ ก็ทำกระบวนการที่ย้อนทิศทาง คือจากความหมายที่ต้องการจะสื่อ ให้หารูปคำที่เหมาะสมสำหรับมโนทัศน์ต่างๆที่เกี่ยวข้อง จากนั้นหาโครงสร้างที่ถูกต้องทางไวยากรณ์สำหรับสร้างประโยคในภาษานั้น และอาจต้องมีการเปลี่ยนรูปคำให้เป็นรูปที่ปรากฏใช้จริงมีวิภัติปัจจัยติดมาด้วย

สำหรับรายละเอียดของแต่ละโมดูลในการประมวลผลภาษานั้น แต่ละระบบก็อาจจะเลือกใช้เทคนิควิธีการและกระบวนทัศน์ที่แตกต่างกันอย่างที่กล่าวมาแล้ว กระบวนทัศน์แบบสถิติเป็นสิ่งที่ระบบประมวลผลภาษาในปัจจุบันใช้กัน ซึ่งเป็นได้ทั้งแบบที่เป็นใช้วิธีการทางสถิติแบบเดียว เช่น การใช้ hidden markove model ในโมดูลของการรู้จำเสียง การใช้ hidden markov model ในการกำกับหมวดคำ เป็นต้น หรืออาจเป็นการใช้สถิติผสมผสานกับการใช้กฎ เช่น การแจงส่วนประโยคโดยใช้ probabilistic context-free grammar

ถึงแม้ว่าแนวทางการใช้สถิติจะเป็นสิ่งที่ยอมรับกันในภาษาศาสตร์คอมพิวเตอร์ปัจจุบัน และถึงแม้ว่า Abney จะพยายามชี้ให้เห็นลักษณะบางอย่างของภาษาที่สามารถอธิบายด้วยวิธีการทางสถิติได้ แต่การใช้สถิติเพียงอย่างเดียวจะเป็นคำตอบของการสร้างแบบจำลองภาษา (language model) ได้จริงหรือ แบบจำลองภาษาจำเป็นต้องมีกฎต่างๆทางภาษาศาสตร์อีกหรือไม่ จำเป็นที่จะต้องมีการผสมผสานระหว่างกฎทางภาษากับการประยุกต์ใช้สถิติหรือไม่ คำถามเหล่านี้เป็นคำถามที่ยังต้องการคำตอบ นักภาษาศาสตร์คอมพิวเตอร์บางกลุ่มก็ไม่สนใจในเรื่องของภาษาศาสตร์อีกเลยเพราะเชื่อว่าสามารถใช้แนวทางทางสถิติเพียงอย่างเดียวในการแก้ปัญหาการประมวลผลภาษาธรรมชาติได้ เช่น ให้คอมพิวเตอร์หาเอาเองว่าหมวดคำที่ควรจะมีในภาษานั้นๆมีอะไรบ้างโดยใช้สถิติเรื่องของการจัดกลุ่ม (clustering analysis) ให้คอมพิวเตอร์หาเอาเองว่าคำแต่ละคำมีความหมายอะไรได้บ้างโดยใช้วิธีการทางสถิติแบบต่างๆ หรือใช้สถิติเพื่อช่วยในการแยกคำ การกำกับหมวดคำและเลือกความหมายของคำหรือค้นหาเอกสารที่ต้องการ (information retrieval) โดยการเปรียบเทียบความสัมพันธ์ระหว่างคำค้นกับคำที่ปรากฏในเอกสารโดยใช้แนวคิดทางสถิติเช่นเวกเตอร์เพื่อคัดเลือกเอกสารที่มีค่าใกล้เคียงสิ่งที่ต้องการมากที่สุด หรือใช้คอมพิวเตอร์เพื่อช่วยวิเคราะห์เอกสารและสรุปสาระสำคัญในเอกสารนั้นโดยวิธีการทางสถิติ ทั้งหมดนี้สามารถกระทำได้โดยให้คอมพิวเตอร์เรียนรู้ลักษณะต่างๆของภาษาจากคลังข้อมูลภาษาจำนวนมหาศาลที่ได้รับ โดยเปรียบเสมือนมนุษย์เรียนรู้ภาษาจากการได้ยินได้ฟังภาษานั้นอยู่เสมอ แน่นอนว่า วิธีการสถิติสามารถช่วยให้คอมพิวเตอร์กระทำสิ่งเหล่านี้ได้อย่างถูกต้องน่าพอใจ แต่สาเหตุที่คอมพิวเตอร์ทำงานได้ถูกต้องเพราะคอมพิวเตอร์ได้เรียนรู้ค่าสถิติที่ต้องการจากข้อมูลภาษาที่ให้ สิ่งเหล่านี้ได้สะท้อนให้เราเห็นหรือไม่ถึงความจริงที่ซ่อนเร้นอยู่ว่าคนเรามีความสามารถทางภาษาได้อย่างไร แบบจำลองของภาษาที่แท้จริงควรจะเป็นอย่างไร เพราะปัญหาเดียวกัน เช่น การกำกับหมวดคำด้วยคอมพิวเตอร์นั้นสามารถกระทำได้โดยใช้วิธีการทางสถิติต่างๆ เช่น ใช้ N-gram model ใช้ maximum entropy ใช้ decision tree คำถามสำคัญจึงอยู่ที่ว่า หากสถิติเกี่ยวข้องโดยตรงกับแบบจำลองภาษาจริง ควรที่จะมีวิธีการทางสถิติแบบเดียวที่แสดงให้เห็นถึงธรรมชาติที่แท้จริงของภาษาหรือไม่ หากว่าควรจะมี สถิติแบบที่เป็นแก่นแท้ของแบบจำลองนี้ควรเป็นอะไร และเราจะตัดสินได้อย่างไร

บรรณานุกรม

Abney, Steven. 1996. Statistical methods and linguistics. In Judith Klavans and Philip Resnik, eds., The Balanacing Act. Cambrdige, MA.:MIT Press. (http://www.sfs.nphil.uni-tuebingen.de/~abney/)

Allen, James. 1995. Natural Language Understanding. 2nd ed. Redwood City: Benjamin/Cummings Publishing.

Backus 1959.

Chomsky, Noam. 1956.

Givon, T. 1995. Functionalism and Grammar. Amsterdam. The Netherlands: John Benjamins Publishing Company.

Levinson, Stephen. 1983. Pragmatics. Cambridge: Cambridge University Press.

Lyons, John. 1977. Semantics. Cambridge: Cambridge University Press.

Jurafsky, Daneil and James H. Martin. 2000. Speech and Language Processing. Englewood Cliffs: New Jersey (Draft)

Manning, Christopher and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

Naur et al (1960)

Hausser, Roland (1999): Foundations of Computational Linguistics. Berlin, New York: Springe (http://www.linguistik.uni-erlangen.de/~rrh/Schriftenverzeichnis.html)

Pinker, Steven. 1994. The language instinct. New York: William Morrow and Company, Inc.