ภาษาศาสตร์คลังข้อมูล

- เป็น theory หรือ methodology?
- ถ้าเป็น theory เป็นการศึกษาใน domain ใด ครอบคลุมเนื้อหาอะไร?
  - ไม่เหมือน morphology, syntax, semantics, etc.
- หรือเป็น purely methodology ที่ใช้ในงานภาษาศาสตร์?
  - ถ้าเป็น ไม่มีผลต่อทฤษฎี เป็นแนวปฏิบัติตามที่กำหนด (given)
  - ทำให้นักทฤษฎีให้ความสำคัญกับข้อมูลมากขึ้น
  - เปลี่ยนโฉมแนวคิดความเข้าใจเกี่ยวกับภาษา
  - การใช้คอมพิวเตอร์และวิธีการทางสถิติมากขึ้น => การเปลี่ยนแนวการศึกษา

คลังข้อมูลภาษา

- คลังข้อมูลภาษาคืออะไร
  - ในปัจจุบัน จะหมายถึงข้อมูลภาษาที่ได้มีการเก็บบันทึกไว้ในระบบคอมพิวเตอร์
  - เป็นข้อมูลที่เป็น representative ของภาษาที่ต้องการศึกษา
  - เก็บตามเงื่อนไขที่กำหนดชัดเจน เพื่อวัตถุประสงค์ที่ต้องการ
  - Corpus ต่างจาก Text archive
- ใช้ประโยชน์อะไร
  - Empirical approach for description of language uses

นิยามของคลังข้อมูลภาษา

- a collection of written or spoken texts (Oxford)
- a large collection of written or spoken language, that is used for studying the language (Longman)
- the collection of a single writer's work or of writing about a particular subject, or a large amount of written and sometimes spoken material collected to show the state of a language (Cambridge)
- A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. (Sinclair, 2005)

คลังข้อมูลภาษา

- คลังข้อมูลภาษาให้ข้อมูลการใช้ภาษาที่เกิดขึ้นจริง จึงเป็นแหล่งข้อมูลที่สำคัญต่อการศึกษาวิจัยต่างๆ เกี่ยวกับภาษา
- คอมพิวเตอร์มีบทบาทสำคัญที่ทำให้สาขานี้เติบโตโดยเร็ว
- แต่การใช้ข้อมูลภาษาที่เกิดขึ้นจริงเพื่อการศึกษาทางภาษามีมาก่อนหน้าแล้ว เริ่มจากการทำเองด้วยมือ
  - 1920s Thorndike & West นับความถี่คำเพื่อเลือกศัพท์ใช้สอน
  - 1930s Zipf พูดถึงความสัมพันธ์ frequency & rank (f x r = c)
  - 1960s Corpus ยุคแรกใช้เครื่อง mainframe

EARLY CORPUS LINGUISTICS

- ก่อนยุค Chomsky งานภาษาศาสตร์อาศัยข้อมูลภาษา
- งานของพวก Structuralist เป็น corpus-based
- งานด้านอื่นๆ ที่ต้องอาศัยข้อมูลภาษา ก็มีมานานแล้ว
  - Language acquisition
  - Language pedagogy
  - Comparative linguistics
  - Lexicography

ยุค CHOMSKY

- มุ่งหา competence ไม่ใช่แค่ describe ภาษา จัดระบบข้อมูลภาษา
- ให้ความสำคัญกับ intuition มากกว่าข้อมูลจริง
- ภาษาเป็น non-finite ไม่มีทางรวบรวมข้อมูลได้ครบถ้วน ต้องอาศัย intuition ข้อมูลเป็น performance มีที่ผิดได้
- Intuition บอกได้ว่าประโยคใด grammatical หรือไม่ จะไม่เห็นประโยค ungrammatical จากข้อมูล (ยกเว้นที่เป็นการพูดผิด)
- ยากที่จะศึกษาข้อมูลขนาดใหญ่ได้ (ปัจจุบัน ไม่ใช่เรื่องยาก)
- เป็นความต่างแนวคิด rationalists กับ empiricists

CORPUS LINGUISTICS

- เห็นว่าข้อมูลจริงเป็นสิ่งจำเป็น
  - ข้อมูลสังเคราะห์บางครั้งก็ตัดสินยากว่า grammatical หรือไม่
  - การแยก grammatical /ungrammatical อาจไม่เหมาะ บางครั้งเป็นเรื่องของ degree การยอมรับ
  - ทฤษฎีที่พัฒนาใช้ได้จำกัด ไม่ครอบคลุมข้อมูลทั่วไป
  - คลังข้อมูลช่วยให้เห็นว่ามีที่เข้าใจคลาดเคลื่อน intuition can be wrong
  - คลังข้อมูลให้สิ่งที่ intuition บอกไม่ได้ เช่น ความถี่ การใช้มากน้อย
- มีทั้งที่เป็น Corpus-based และ corpus-driven

ตัวอย่าง CONCORDANCE OUTPUT

taste it is that such poor cattle always have in their mouths

of sparing the poor child the inheritance of any part

small property of my poor father, whom I never saw--so long

desolate, while your poor heart pined away, weep for it

Miss, if the poor lady had suffered so intensely

the love of my poor mother hid his torture from me

stockings, and all his poor tatters of clothes, had, in a long

faded away into a poor weak stain. So sunken and

on your way to the poor wronged gentleman, and, with a

detachment from the poor young lady, by laying a brawny

Concordance = an alphabetical list of all the words used in a book or set of books, with information about where they can be found and usually about how they are used

พัฒนาการคลังข้อมูลภาษา

- ยุคแรก
  - 1960 Quirk สร้าง “Survey of English Usage” (SEU) (British English) มีทั้งภาษาพูดและภาษาเขียนที่ใช้ในระหว่างปี 1953-1987 ขนาด 1 ล้านคำ โดยแบ่งเป็นข้อมูลภาษาพูดและภาษาเขียนอย่างละครึ่งหนึ่ง
  - 1961 Francis และ Kucera สร้าง Brown Corpus (American English) เก็บจากสิ่งพิมพ์ทั่วไปในปีนั้น มีขนาด 1 ล้านคำ เสร็จปี 1964
  - 1975 Svartvik สร้าง London-Lund corpus (LLC) โดยนำข้อมูลภาษาพูดจาก SEU มาเก็บในคอมพิวเตอร์ และเพิ่มจำนวนเป็น 1 ล้านคำ

พัฒนาการคลังข้อมูลภาษา

- ยุคแรก
  - Lancaster-Oslo/Bergen (LOB) เป็นคลังข้อมูลคู่เทียบของคลังข้อมูล Brown เป็น British English ปี 1961, จัดสร้างขึ้นในช่วงปีค.ศ. 1970-1978
- ยุคต่อมา
  - Birmingham Collection of English Language Corpus ของ Sinclair มีขนาด 20 ล้านคำ (British English) ต่อมาขยายขึ้นเรื่อยๆ เรียก Bank of English ในปี ค.ศ. 2005 คลังข้อมูลนี้ได้มีขนาดถึง 450 ล้านคำ

พัฒนาการคลังข้อมูลภาษา

- ยุคต่อมา
  - Longman Corpus Network ประกอบด้วย Longman/Lancaster English Language Corpus (LLELC), Longman Spoken Corpus (LSC) และ Longman Corpus of Learners' English (LCLE) แต่ละคลังข้อมูลมีโครงสร้างและเนื้อหาข้อมูลที่แตกต่างกัน เพื่อวัตถุประสงค์ที่แตกต่างกัน คลังข้อมูลทั้งสามนี้ใช้เพื่อประโยชน์ในการทำพจนานุกรม
  - British National Corpus มีขนาด 100 ล้านคำ ภาษาเขียน 90% พูด 10%

การแยกประเภทคลังข้อมูล

- General vs Specialized corpora
- Written vs Spoken corpora
- Native vs Learner corpora
- Plain text vs annotated corpora
- Mono vs Multilingual corpora
- Parallel corpora vs Comparable corpora
- Static vs Monitor

ต่างมุมมองจากผู้ที่เกี่ยวข้อง

- คนสร้างคลังข้อมูล : encoding
- คนพัฒนาโปรแกรมคอมพิวเตอร์ : software tools ทางภาษา
- คนใช้คลังข้อมูลเพื่องานภาษาศาสตร์ : empirical, corpus-based, probabilistic
- คนใช้คลังข้อมูลในงานประยุกต์ต่างๆ
  - การสอนภาษา DLL, lexical approach, phraseology
  - การแปล Parallel corpus, Translation study
  - การทำพจนานุกรม ประมวลศัพท์ : corpus-based
  - การประมวลผลภาษา : training