Corpus Design

- วัตถุประสงค์ ต้องการ general corpus หรือ specialized corpus
- ขนาดของ corpus
  - 1 ล้านคำ? 10 ล้านคำ? 100 ล้านคำ?
  - Sample ขนาดเท่าๆ กัน? ผลของการ fix ขนาด sample text?
- corpus เป็น static หรือ dynamic (monitor corpus)
  - เก็บแบบ opportunistic หรือ plan
- representative and balance
  - balance หมายถึงมีจำนวนเท่าๆกัน?
  - ใช้เกณฑ์อะไรในการเลือก text? จำนวนผู้อ่าน? ความสำคัญ? เวลา? ผู้เขียน?

Corpus Design

- คนที่สร้าง corpus ควรเป็น expert in the communicative patterns of the communities คือ เป็นคนที่ใช้และรู้จักภาษาที่ต้องการศึกษา
- ไม่ควรเป็น expert in corpus analysis เพราะความที่รู้ว่ามีอะไรใน corpus และต้องการอะไรจากการใช้ corpus อาจมีผลทำให้เลือกเฉพาะข้อมูลที่มีสิ่งที่ตนเองต้องการ
- เลือก text จาก external criteria ไม่ใช่ internal criteria

=> 1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.

Corpus Design

- เราสร้าง Corpus ก็เพื่อใช้กับวัตถุประสงค์ที่ต้องการ
- Corpus ต้องเป็นตัวแทนของภาษาที่ต้องการศึกษา
- ภาษาจริงนั้นมีไม่จำกัด, corpus เป็นเพียงการสุ่มตัวอย่าง จึงไม่มีทางจะมี exact characteristics เหมือนภาษาจริงได้ ไม่มีแม้แต่ correct proportion

=> 2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

Corpus Design

- เลือกข้อมูลภาษาอย่างไรให้เป็น representative?
- คำนึงถึงสามเรื่อง 1. แนวภาษาที่ต้องการ
  2 เกณฑ์ในการเลือกตัวอย่าง, 3. มิติต่างๆของการสุ่มตัวอย่าง
- **1. แนวภาษาที่ต้องการ (orientation)** เป็นตัวกำหนดข้อมูลที่เลือก
  Brown corpus ต้องการเก็บภาษามาตรฐาน เลือกแต่งานที่ตีพิมพ์ ไม่มีความต่างภายในมาก
- historical corpus ต้องการข้อมูลที่มีความต่างภายในว่ามีข้อมูลภาษาจากช่วงเวลาต่างๆ อยู่
- parallel corpus ต้องการข้อมูลที่มีความต่างหลายภาษาที่เทียบกันได้

⇒3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.

Corpus Design

⇒3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.

- การใช้ข้อมูลใน corpus ต้องระวังถ้าจะนำเฉพาะบางส่วนมาเทียบกัน ต้องแน่ใจว่าได้ถูกออกแบบมาให้เทียบกันได้
- ด้วย software ปัจจุบัน เราสามารถเลือกเฉพาะส่วนที่ต้องการได้ "dial-a-corpus" แต่ต้องระวังว่าการใช้เฉพาะส่วน represent ภาษาอย่างที่ต้องการจริงหรือไม่ เพราะ corpus นั้นอาจมี variety ของ text ต่างๆเพียงพอที่จะเป็น normative corpus แต่ว่าแต่ละส่วนไม่เพียงพอเป็นตัวแทนของแต่ละ variety นั้น

Corpus Design

- **2. เกณฑ์ที่ใช้ในการเลือก** text **เป็นสิ่งสำคัญ มีหลากหลาย**
  - Mode : spoken, written
  - Text type : book, journal, letter, memo, …
  - Domain : mathematics, physics, arts, …
  - Location : british, australia, …
  - Date : old, middle, …
- Corpus **ที่ต้องการจะเป็นตัวกำหนดเกณฑ์การเลือกข้อมูลในตัว เช่น** MICASE (Michigan Corpus of Academic Spoken English) => spoken, academic, american, Michigan

Corpus Design

- ผู้สร้าง corpus ควรเลือก criteria ที่ชัดเจน ตัดสินง่าย เพื่อเลี่ยงปัญหาในการนำข้อมูลเข้า เพราะถ้ามีข้อสงสัยเรื่องข้อมูลขึ้นมา corpus ไม่ว่าจะมีขนาดใหญ่เพียงใด ก็จะขาดความน่าเชื่อถือ

=> 4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.

Corpus Design

- นอกจาก criteria หลัก information อื่นก็สามารถนำเข้าได้ เช่น gender, age, demographic, … เพื่อประโยชน์ในการพิจารณาข้อมูล ซึ่งผู้อื่นอาจต้องการเลือกเฉพาะส่วนภายหลัง แต่ควรแยกจากข้อมูลภาษาให้ชัดเจน

=> 5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.

Corpus Design

- **3. การสุ่มตัวอย่าง**
  - Criteria หลักๆ จะเป็นตัวแยกข้อมูลเป็นส่วนๆ หากมอง intersection ของหลาย criteria จะมองเทียบได้เป็น cell
  - เช่น การสร้าง spoken corpus ที่มี mode = private, public participant : 3 คนขึ้นไป หรือ น้อยกว่า 3 คน, ได้เป็น 2 x 2 cell
  - แต่ละ criteria จึงเป็นการแบ่ง cell ย่อยไปเรื่อยๆ
  - ให้คิดต่อว่าใน cell หนึ่งๆ ต้องการข้อมูลน้อยสุดเท่าใดสำหรับงานที่ต้องการศึกษา
  - ขนาด corpus ประมาณจาก จำนวน cell x minimum size
  - ในความเป็นจริง ต้องดูว่า แต่ละ cell นั้นเป็นไปได้ มีเกิดจริง เก็บข้อมูลได้

Corpus Design

- การเลือกบางส่วนจาก text ยาว แต่ละส่วนก็มีความต่างกัน ไม่สามารถ assume ว่า represent text ทั้งหมดได้
- การนำ text ยาวทั้งหมดลง corpus ก็ต้องไม่ให้หนัก corpus ไป อาจบรรเทาปัญหานี้ได้โดยการสร้าง corpus ใหญ่มากๆ
- แต่การนำ text ทั้งหมดก็มักมีปัญหากับเจ้าของลิขสิทธิ์
- การสุ่ม text ด้วย sample size เท่ากันหมด ก็ไม่ใช่วิธีที่ทำกันในปัจจุบัน เพราะไม่มีเหตุผลทางภาษาศาสตร์

=> 6. Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

Corpus Design

- Representative ดูจากผู้พูด/ใช้ภาษาที่เราต้องการว่า เขาเขียน/อ่าน text อะไร
- จะเลือกจากสัดส่วน publication, จากความสนใจอย่างไร
- จะเลี่ยงไม่เอาแต่ text ที่ได้โดยสะดวกอย่างไร เช่น จาก web, public domain, …
- ถ้าคนอ่านไทยรัฐมากที่สุด ต้องเอา text จากไทยรัฐมากกว่านสพ.อื่น? ภาษาไทยในไทยรัฐเป็นภาษาที่เหมาะ? จะใช้ prescriptive view?
- สิ่งแรกคือ กำหนด criteria ที่บอกโครงสร้าง corpus และใช้กำหนด component ของ corpus
- แต่ละ component ดูว่าควรนำ text type อะไรเข้าโดยใช้ external criteria

Corpus Design

- จัด priority ของ text type โดยดู factor ต่างๆ ที่เกี่ยวข้องกับแต่ละ text type
- กำหนดขนาดที่ต้องการของแต่ละ text type จำนวนรวม จำนวน text หนทางในการรวบรวม text
- ระหว่างทำ ให้เทียบที่วางแผนไว้กับที่รวบรวมได้จริง
- บันทึกปัญหาและสิ่งที่ทำในระหว่างโครงการ เพื่อช่วยสรุปภาพรวม corpus สุดท้ายที่ได้
- ตย.การสร้าง Bank of English ใน 1980s ต้องการรวมนิยายดีๆ เพราะคิดว่าเป็นตัวอย่างงานเขียนคุณภาพ เมื่อนำไปใช้การสอนภาษา พบว่าเราไม่ต้องการตัวอย่างการใช้คำหรูๆเหล่านั้น การบันทึกการทำงานทำให้แก้ไขและจัดระบบข้อมูลให้เหมาะภายหลังได้ โดยเพิ่มข้อมูลอื่นให้สมดุล

Corpus Design

- จำเป็นต้องให้ผู้ใช้รู้รายละเอียด corpus มากสุด เพื่อตีความผลที่พบได้ว่าเป็นเพราะ text ที่คัดเลือกมาหรือไม่

=> 7. The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

- Balance เป็น concept ที่ vague ยิ่งกว่า representative
- Corpus อาจเรียกว่า balance ได้ถ้าสัดส่วน text ต่างๆสอดคล้องกับ intuitive judgment
- General corpus ส่วนใหญ่ไม่ balance เพราะมีส่วน spoken ไม่ถึงแม้แต่ 50% สัดส่วนจริงอาจมากถึง 90% ก็ได้
- specialized text ใน general corpus จะเลือกอย่างไรจึง balance

Corpus Design

- ตย. Popular magazine มีมากมาย ถ้าเลือก magazine แค่ คอมพิวเตอร์, กีฬา, ดนตรี อาจไม่ balance พอ

=> 8. The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.

- มีการใช้ topic เป็นตัวเลือก text เพื่อสร้าง corpus
- แต่การใช้ topic เป็นเกณฑ์การเลือก text เป็นการใช้ internal criteria จึงไม่ควรทำ (topic กำหนด vocab ซึ่งเป็นเรื่องของ text)
- สามารถใช้ external criteria มาเป็นตัวเลือกแทน topic และก็จะมีผลต่อ vocab ที่ได้ในทางอ้อม เช่น text ใช้ในวิชาชีพ ในการศึกษา

Corpus Design

- => 9. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.
- เรื่องขนาด corpus เท่าไรนั้น ขึ้นกับงานที่ต้องการใช้ คำถาม และวิธีการหาคำตอบ ไม่มี maximum size สำหรับ corpus
- ใน 1 ล้านคำ ของ Brown corpus มี 69,002 ศัพท์, 35,065 ศัพท์พบเพียง 1 ครั้ง
- เป็นลักษณะปกติของภาษา สอดคล้อง Zipf's law (Freq ~ 1/Rank, Freq * Rank = k) อาจพบว่าคำเกิด 1 ครั้งมีครึ่งของข้อมูล, คำเกิด 2 ครั้งมีหนึ่งในสี่, ...
- การกระจายตัวแบบนี้พบในคำหลาย category, หลายความหมายด้วย

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table 1.3 Empirical evaluation of Zipf's law on *Tom Sawyer*.

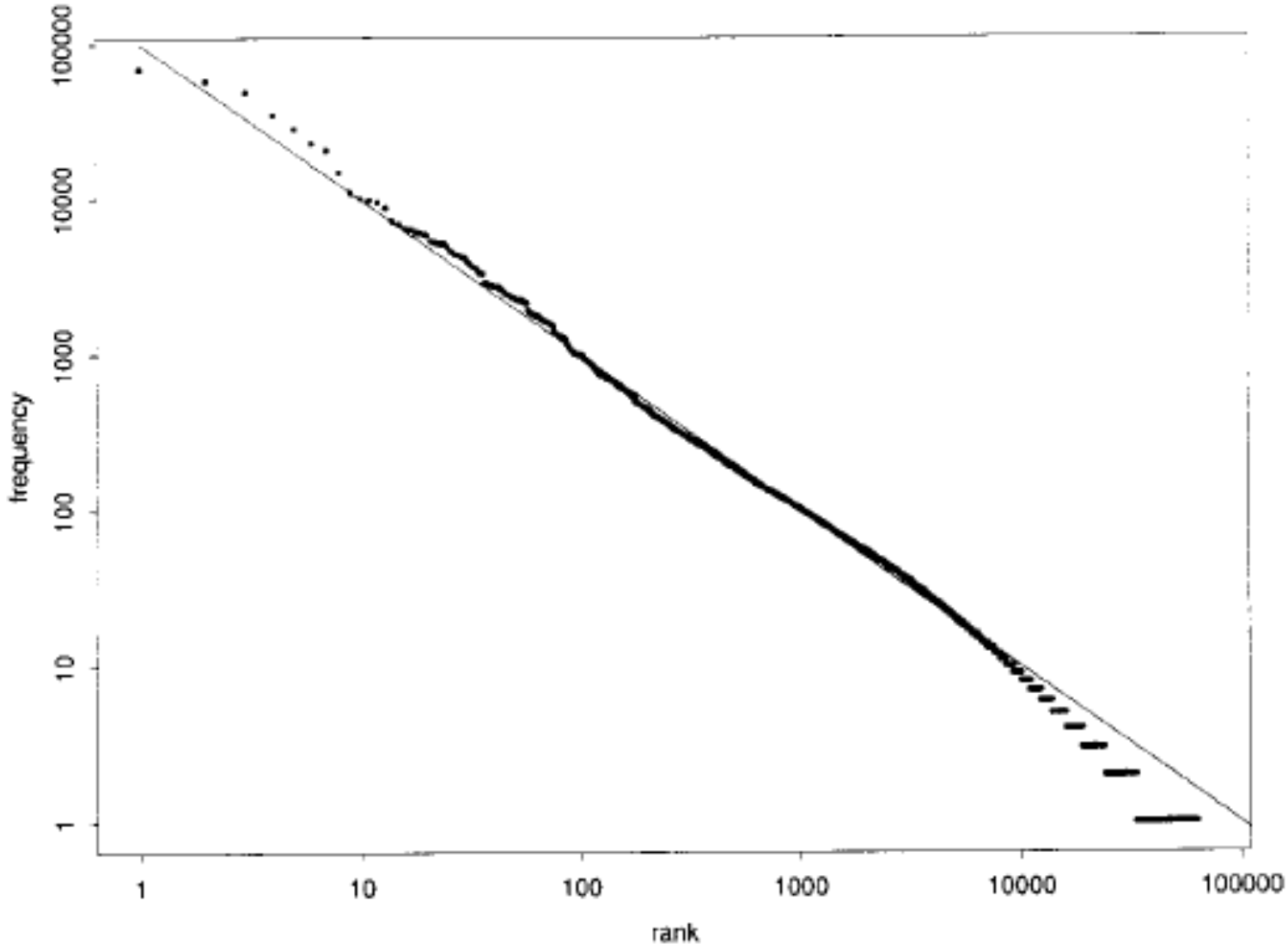


Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for $k = 100,000$, that is $f \times r = 100,000$.

Corpus Design

- ขนาด corpus ขึ้นกับคำที่ต้องการศึกษา ถ้าคำไม่กำกวม ควรได้อย่างน้อย 20 ตย. ถ้าคำมีความกำกวม ก็ต้องการ 20 + 20/2 + 20/3 + 20/4 + … อาจถึง 50 ตย.

Rank	Freq	K=Rank*Freq
1	20	20
2	10	20
3	6.7	20
4	5	20
5	4	20
6	3.3	20
7	2.9	20
8	2.5	20
9	2.2	20
10	2	20

Corpus Design

	Pr(X)	Pr(Y)		Pr(X-Y)
1	0.00002	0.00002	1	4E-10
1,000,000	20	20	50,000,000,000	20

- ถ้าสนใจมากกว่าคำเดียว เช่น ดู compound 2 คำ สมมติแต่ละคำพบ 20 ครั้งใน 1 ล้านคำ ถ้าจะพบ 2 คำนั้นด้วยกัน 20 ครั้ง ข้อมูลควรมี 5 หมื่นล้านคำตามหลักคณิตศาสตร์ แต่ความจริงน้อยกว่านั้น อาจใช้ 25 ล้านคำก็พอ
- อาจประมาณขนาดจากการวิเคราะห์เบื้องต้นก่อน ดูว่าได้เท่าไร และจะเพิ่มขนาด corpus เท่าใดจึงจะได้ตย.พอสำหรับการวิเคราะห์
- ใน specialized corpus ขนาดของ corpus จะน้อยกว่า general corpus ได้มาก เพราะจำนวนศัพท์จะน้อยกว่า จำกัดเฉพาะในเรื่อง

Corpus Design

- HK corpus ด้าน comp sci. เทียบกับ LOB 1 ล้านคำเท่ากัน
- Specialized corpus มีจำนวนศัพท์น้อยกว่า เกิดซ้ำๆ มากกว่า

	LOB	HK	%
Number of different word-forms (types)	69990	27210	39%
Number that occur once only	36796	11430	31%
Number that occur twice only	9890	3837	39%
Twenty times or more	4750	3811	80%
200 times or more	471	687	(69%)

Table 1. Comparison of frequencies in a general and a specialised corpus.

Corpus Design

- ความต่างที่เห็นเป็นเรื่องของ homogeneity
- Specialized corpus มีลักษณะที่เป็น homogeneity มากกว่า
- Homogeneity เป็น criteria เลือก text เข้า corpus ได้ โดยการดูว่า text ไหนที่แปลกแยกกว่าก็ละไปไม่เอาเข้า แต่หากพบ text แบบนั้นมากๆเข้า แสดงว่าการวางโครงสร้าง corpus อาจไม่ถูก มีปัญหา เพราะมีการรวบ distinct text type 2 อย่างเข้าด้วยกัน

=> 10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

Corpus Design

- What is not a corpus
  - World wide web ไม่มี dimension ชัดเจน เปลี่ยนตลอดเวลา
  - Archive มีวัตถุประสงค์เพื่อเก็บสะสม text
  - Collection of citation เป็นการคัดข้อมูลเล็กๆ ไม่มีความต่อเนื่องตัวบท
- **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**

http://www.natcorp.ox.ac.uk/

MICASE

Latest Headlines | Nation | Krungtheptura... | WebMail – CU | Gmail | Facebook | Linguistics | User Identification | Bookmarks

BRITISH NATIONAL CORPUS

About

What is the BNC?
Creating the BNC
BNC Products
Copyright
Contact Us
Contents A-Z

Using the BNC

What can I do with the BNC?
Using BNC with Xaira
FAQ

Obtaining

How to order
Pricing
Xaira
FAQ

About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [more]

Search the Corpus

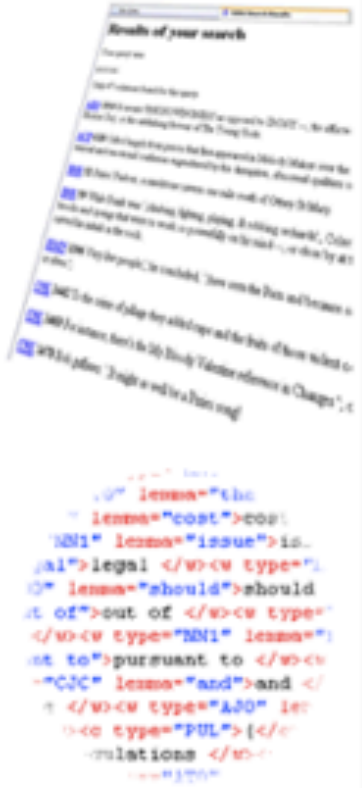
Type a word or phrase in the search box and press the Go button to see up to 50 random hits from the corpus.

Look up: Enter text Go

You can search for a single word or a phrase, restrict searches by part of speech, search in parts of the corpus only, and much more.

The search result will show the total frequency in the corpus and up to 50 examples. [more information]

News from the BNC



British National Corpus

- British English ภาษาปัจจุบัน
- เป็น general purpose corpus ขนาด 100 ล้านคำ
- เป็นภาษาเขียน 90% ภาษาพูด 10%
- POS tagged โดยโปรแกรม CLAW
- encoding ตามมาตรฐานที่กำหนดโดย TEI
- core corpus = 2 ล้านคำ มีภาษาเขียน = ภาษาพูด = 50% ตรวจ POS tag ให้ถูกต้อง

British National Corpus

- The British National Corpus is:
  - a sample corpus: composed of text samples generally no longer than 45,000 words.
  - a synchronic corpus: the corpus includes imaginative texts from 1960, informative texts from 1975.
  - a general corpus: not specifically restricted to any particular subject field, register or genre.
  - a monolingual British English corpus: it comprises text samples which are sub-stantially the product of speakers of British English.
  - a mixed corpus: it contains examples of both spoken and written language

British National Corpus

- BNC Consortium led by Oxford University Press, other members are Addison-Wesley Longman and Larousse Kingfisher Chambers; academic research centres at Oxford University Computing Services (OUCS), the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, and the British Library's Research and Innovation Centre.
- The project was funded by the commercial partners, the Science and Engineering Council (now EPSRC) and the DTI under the Joint Framework for Information Technology (JFIT) programme.
- โครงการจัดทำ BNC เริ่มในปี 1991 และมี first release ในปี 1995

British National Corpus

- planning stage : the design principles were drawn up. These principles included the selection criteria that were used as the basis for the collection of the texts
- Once a suitable texts was identified and permission to use it had been obtained, the text was converted to machine readable form
- The text was then passed to UCREL, where word class tagging was automatically added.
- Correction and validation of the bibliographic and contextual information in all the BNC Headers was also carried out for this second version of the corpus, known as the BNC World Edition. BNC World was made available for world-wide distribution in 2001

British National Corpus

Table 1. Composition of the BNC World Edition

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08
Spoken context-governed	757	6135671	6.28	428558	7.07
All Spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All Written	3144	87278205	89.39	5013972	82.82

Table 2. Date of production

Creation date	*texts*	*w-units*	*%*	*s-units*	*%*
Unknown	162	1814051	1.85	127132	2.10
Before 1974	47	1741624	1.78	121323	2.00
1974 to 1983	156	4621950	4.73	255057	4.21
1984 to 1994	3689	89442309	91.62	5549581	91.68

British National Corpus

- Design ของ written text
- การเลือก text พิจารณาทั้งด้านการสร้างสาร และการรับสาร
- text ที่ตีพิมพ์ไม่ใช่ตัวแทนทั้งหมดของ written text
- There was no single source of information about published material that could provide a satisfactory basis for a sampling frame ต้องใช้ข้อมูลหลายแหล่ง
- Catalogues of books published per annum บอกเกี่ยวกับ production ไม่บอกว่ามีการอ่านมากหรือน้อย

British National Corpus

- books in print บอกข้อมูลหนังสือที่ยังตีพิมพ์ แสดงถึงหนังสือส่วนหนึ่งว่ายังเป็นที่สนใจอ่าน
- best seller list แสดงส่วนที่เป็นที่นิยมอ่าน
- สถิติการยืมในห้องสมุดแสดงส่วนของการอ่าน
- written text เลือกจาก 3 เกณฑ์หลัก : domain, time, medium
  - domain : 75% เป็น informative 25% เป็น imaginative
  - medium : 60% เป็นหนังสือ 30% เป็นวารสาร หนังสือพิพม์ 10% จากแหล่งอื่นๆ
  - time : ใช้ตั้งแต่ปี 1975 ขึ้นไป ยกเว้น imaginative text บางอันที่เป็นนิยมอ่านอยู่

British National Corpus

- text sample มา 40,000 คำ ถ้า text นั้นสั้นกว่า 40,000 คำจะตัดออก 10% เพื่อกันปัญหาลิขสิทธิ์ บาง text ก็เอามา >40,000 แต่ไม่เกิน 45,000 คำ sample เอามาแบบสุ่มตอนต้น ตอนกลาง หรือตอนท้าย
- ประมาณครึ่งของ text จาก book เลือกจากรายการใน book in print แบบสุ่ม และคัดเฉพาะที่เป็น british author และตรงกับ criteria ที่วางไว้
- ส่วนที่เหลือเลือกอย่างเป็นระบบให้ text ทั้งหมดได้ตามแบบที่วางไว้

British National Corpus

Table 3. Written domain

Domain	*texts*	*w-units*	*%*	*s-units*	*%*
Applied science	370	7104635	8.14	357067	7.12
Arts	261	6520634	7.47	321442	6.41
Belief and thought	146	3007244	3.44	151418	3.01
Commerce and finance	295	7257542	8.31	382717	7.63
Imaginative	477	16377726	18.76	1356458	27.05
Leisure	438	12187946	13.96	760722	15.17
Natural and pure science	146	3784273	4.33	183466	3.65
Social science	527	13906182	15.93	700122	13.96
World affairs	484	17132023	19.62	800560	15.96

Table 4. Written medium

Medium	*texts*	*w-units*	*%*	*s-units*	*%*
Book	1414	49891770	57.16	2895652	57.75
Periodical	1208	28356005	32.48	1487725	29.67
Published miscellanea	238	4197450	4.80	288004	5.74
Unpublished miscellanea	249	3508500	4.01	222438	4.43
To-be-spoken	35	1324480	1.51	120153	2.39

British National Corpus

- ดูจาก best seller คัดรายการที่ซ้ำกับที่เลือกแบบสุ่มไปแล้วออก คัดที่ไม่ใช่ british author ออก กันไม่ให้มี text ของ author คนเดียวกันเกิน 120,000 คำ นอกจากนี้ยังดูจาก การได้รางวัล การยืมจากห้องสมุดสาธารณะ การยืมข้ามห้องสมุด
- ก่อนจะนำ text มาใช้ ต้องติดต่อส่งแบบฟอร์มขออนุญาติจากเจ้าของลิขสิทธิ์ text ที่ไม่ได้รับคำตอบรับต้องกันออกจากรายการ

British National Corpus

- Design of Spoken text
- แยกออกเป็นสองส่วนเท่าๆกัน คือ ส่วนที่เป็นภาษาพูดแยกตาม demographic ซึ่งเป็นส่วนที่เป็นการถอดเสียงของบทสนทนาต่างๆของอาสาสมัครจากภูมิลำเนาต่างๆ และส่วนที่เป็นภาษาพูดแยกตาม context-governed ซึ่งเป็นส่วนที่มาจากการถอดเสียงของเทปบันทึกการประชุมต่างๆ ที่ได้จากสถานการณ์แบบต่างๆ
- ข้อมูลที่เป็น demographic ได้มาจากอาสาสมัคร 124 คนจากกลุ่มสังคมต่างๆ ทั้งชาย และหญิงอายุต่างๆ อาสาสมัครมาจากภูมิลำเนา 38 แห่งทั่วสหราชอาณาจักร

British National Corpus

- การคัดเลือกพยายามให้มีเพศชายหญิงเท่ากัน มีจำนวนคนในแต่ละกลุ่มอายุเท่าๆกัน และในแต่ละกลุ่มสังคมเท่าๆกัน โดยที่อาสาสมัครจะอัดเทปการสนทนาของตัวเองที่เกิดขึ้นในชีวิตประจำวันเป็นเวลา 2-3 วัน พร้อมทั้งจดบันทึกของบทสนทนาแต่ละบทไว้ว่าเกิดขึ้นที่ไหนกับใครและเมื่อไร
- ไม่ได้ใช้คนเป็น 1,000 เพราะจะยุ่งยากในการจัดการ จึงใช้ร้อยกว่าคนแต่พยายามให้ครอบคลุมความหลากหลาย

British National Corpus

- ข้อมูลที่เป็น context-governed ตั้งเป้าหมายว่าจะเก็บคำพูดต่อเนื่องที่บันทึกไว้ในสี่สาขาในปริมาณเท่าๆกัน ดังต่อไปนี้
- 1. ด้านการศึกษาหรือการให้ข่าวสารข้อมูล เช่น การบรรยายในห้องเรียน การรายงานข่าว การอภิปรายในชั้นเรียน การอบรม
- 2. ด้านธุรกิจ เช่น การสาธิตประกอบการขาย การประชุมการค้า การให้คำปรึกษา การสัมภาษณ์
- 3. ด้านองค์กรและสาธารณะ เช่น การกล่าวสวด การกล่าวปราศรัย การประชุมกรรมการ การประชุมสภา
- 4. ด้านบันเทิง เช่น การบรรยายกีฬา การประชุมในคลับ การสนทนาผ่านรายการวิทยุ

British National Corpus

In addition, the following classifications are applicable to both demographic and context-governed spoken texts:

Table 17. Region where spoken text captured

Region	*texts*	*w-units*	*%*	*s-units*	*%*
Unknown	35	446584	4.31	27706	2.66
South	312	4658232	45.04	458253	44.10
Midlands	213	2471184	23.89	240320	23.12
North	350	2765729	26.74	312842	30.10

Table 18. Interaction type for spoken text

Interaction type	*texts*	*w-units*	*%*	*s-units*	*%*
Monologue	212	1578614	15.26	94272	9.07
Dialogue	698	8763115	84.73	944849	90.92

open data for language research and education

Download
oanc masc other

Contribute
texts annotations
derived data

The Open American National Corpus

The Open American National Corpus (OANC) is a massive electronic collection of American English, including texts of all genres and transcripts of spoken data produced from 1990 onward. All data and annotations are fully open and unrestricted for any use.

Available Data and Annotations

OANC : 15 million words of contemporary American English with automatically-produced annotations for a variety of linguistic phenomena.

What's New

Penn Treebank Syntax

Penn Treebank Syntax: syntax annotations for the entire 500K words of MASC in the original PTB (bracketed) format.

ColnCo

American National Corpus

- BNC เป็นข้อมูล British English ต้องการ American English
- Linguistic Data Consortium (LDC) ใน US distribute corpus แต่ส่วนใหญ่เป็น corpus เฉพาะ เก็บข้อมูลสะดวก ไม่มีปัญหาลิขสิทธิ์
- ต้องการสร้างเพื่อใช้ในงานต่างๆ เช่น computational linguistics, lexicography, speech recognition and synthesis, literary studies, and all varieties of linguistics.
- 100 ล้านคำ เป็น general corpus และ comparable กับ BNC
- กำกับข้อมูลตามมาตรฐาน Corpus Encoding Standard

American National Corpus

- Proposed ในปี 1998
- the ANC project is undertaken in cooperation with a consortium of publishers, organizations, and academic institutions in the US.
- in October of 2003 the first 11.5 million words of the ANC were released, second release = 22 million words
- corpus of 100 million words of American written and spoken language that generally follows the framework of the BNC
- The ANC will only contain texts from 1990 on, while the BNC contains texts from 1960 – 1993.

American National Corpus

- The ANC, however, will contain electronic texts such as e-mail, webpages, and e-talk from chat rooms.
- Core ANC เป็นส่วนที่เหมือน BNC
- Satellite ANC เป็น specialized corpora เพิ่มเข้ามาจากความร่วมมือโครงการอื่น เช่น ICE, MORE, LAWS
- The ANC is encoded in XML and is conformant to the XML Corpus Encoding Standard (XCES) schemas for primary data and annotations.

American National Corpus

- linguistic annotations are contained in separate XML documents linked to the original rather than being interspersed with the original data in a single XML document.
- Part of speech annotation of the ANC has been done using the Biber tagger. The ANC is also being tagged with the C5 and C7 versions of the CLAWS tagger
- Some of the major challenges of creating the ANC are selection and acquisition of texts; legal issues related to copyright and use of the texts; and transduction of the texts into a common format

American National Corpus

- Selecting -> acquiring -> copy right agreement
- The First and Second Releases of the ANC include materials which have been acquired to date, and therefore the current release of the ANC is not balanced
- provides an opportunity to identify bugs and user issues
- The CD containing the second release of the ANC can be ordered from the LDC.

55% Books

Non-fiction 41%

Natural Sciences (e.g., Biology Chemistry Geology Math, Physics)
Applied Sciences (e.g., Communication, Energy, Engineering)
Social Sciences (e.g., Anthropology, Geography, Psychology Sociology)
World Affairs (e.g., Economics, Government, History, Politics)
Commerce (e.g., Business, Finance, Industry, Economics)
Arts (e.g., Media, Performing Arts, Visual Arts)
Beliefs (e.g., Mythology, Astrology, Philosophy Religion)
Leisure (Antiques, Gardening, Hobbies, Travel)
Biographies

Fiction 14%

General fiction
Historical fiction
Science fiction
Romantic fiction
Mystery
Adventure
Poetry
Drama
Humor

20% Newspapers, Magazines and Journals

Magazines – General & Specialized

Newspapers – National and Local (Daily and Weekly)

Journals

10% Spoken

Face-to-face

Meetings

Phone conversations

Planned speeches

10% Electronic

Web pages; E-mail; Chat rooms

5% Miscellaneous (published & unpublished material)

Solicitation letters; Brochures; Memos

Figure 1: Target Design of the 100 million word American National Corpus (ANC).

American National Corpus

- ANC consortium : Pearson Education,· Random House Publishers,· Langenscheidt Publishing Group, Harper Collins Publishers, Cambridge University Press, LexiQuest, Microsoft Corporation, Shogakukan,Inc. Associated Liberal Creators Press, Taishukan Publishers, Oxford University Press, Kenkyusha Publishers, International Business Machines Corporation
- สมาชิกใน consortium ร่วมออกค่าใช้จ่ายแต่ได้สิทธิ์การใช้ก่อน
- LDC ช่วยเรื่องการขอลิขสิทธิ์

ANC

Text type	Text name	No. of texts	No. of words	Contributor
Spoken	Callhome	24	50,494	LDC
Spoken	Switchboard	2320	3,056,062	LDC
Spoken	Charlotte Narrative	95	117,832	Project MORE
			TOTAL SPOKEN	3,224,388
Written	New York Times	4148	3,207,272	LDC
Written	Berlitz Travel Guides	101	514,021	Langensheidt Publishers
Written	Slate Magazine	4694	4,338,498	Microsoft
Written	Various non-fiction	27	224,037	Oxford University Press
			TOTAL WRITTEN	8,283,828
			TOTAL CORPUS	11,508,216

Table 1. Composition of the ANC First Release

Data

Second release

Spoken			
Corpora	**Domain**	**No. files**	**No. words**
callhome	telephone	24	52,532
charlotte	face to face	93	198,295
micase	academic discourse	50	593,288
switchboard	telephone	2,307	3,019,477
Spoken Totals		2,474	3,863,592
Written			
Corpora	**Domain**	**No. files**	**No. words**
911 report	government, technical	17	281,093
berlitz	travel guides	179	1,012,496
biomed	technical	837	3,349,714
buffy	blog	143	3,093,075
hargraves	fiction	106	405,195
eggan	fiction	1	61,746
icic	letters	245	91,318
nytimes	newspaper	4,148	3,625,687
oup	non-fiction	45	330,524
plos	technical	252	409,280
slate	journal	4,531	4,238,808
verbatim	journal	32	582,384
web data	government	285	1,048,792
Written Totals		10,821	18,530,112
Corpus Totals		13,295	22,393,704

Open ANC

Contents

Spoken			
Name	**Domain**	**No. files**	**No. words**
charlotte	face to face	93	198,295
switchboard	telephone	2,307	3,019,477
Spoken Totals		2,410	3,217,772

Written			
Name	**Domain**	**No. files**	**No. words**
911 report	government, technical	17	281,093
berlitz	travel guides	179	1,012,496
biomed	technical	837	3,349,714
eggan	fiction	1	61,746
icic	letters	245	91,318
oup	non-fiction	45	330,524
plos	technical	252	409,280
slate	journal	4,531	4,238,808
verbatim	journal	32	582,384
web data	government	285	1,048,792
Written Totals		6424	11,406,155
Corpus Totals		8,832	14,623,927

INTERNATIONAL CORPUS OF ENGLISH

ICE Teams:

Bahamas

GO!

The International Corpus of English (ICE) began in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Twenty-three research teams around the world are preparing electronic corpora of their own national or regional variety of English. Each ICE corpus consists of one million words of spoken and written English produced after 1989. For most participating countries, the ICE project is stimulating the first systematic investigation of the national variety. To ensure compatibility among the component corpora, each team is following a common corpus design, as well as a common scheme for grammatical annotation.

News April 2011

Tagging ICE Corpora

Currently available ICE corpora:

- Canada*
- East Africa*
- Great Britain
- Hong Kong*
- India*
- Ireland
- Jamaica*
- New Zealand
- The Philippines*

International Corpus of English

- ICE began in 1990 with the primary aim of collecting material for comparative studies of English worldwide.
- Eighteen research teams around the world are preparing electronic corpora of their own national or regional variety of English
- Each ICE corpus consists of one million words of spoken and written English produced after 1989.
- To ensure compatibility among the component corpora, each team is following a common corpus design, as well as a common scheme for grammatical annotation.

International Corpus of English

- Each component corpus contains 500 texts of approximately 2,000 words each - a total of approximately one million words.
- The texts in the corpus date from 1990 or later. The authors and speakers of the texts are aged 18 or over, were educated through the medium of English, and were either born in the country in whose corpus they are included, or moved there at an early age and received their education through the medium of English in the country concerned.

ICE Text Categories

Numbers in brackets indicate the number of 2,000-word texts in each category.

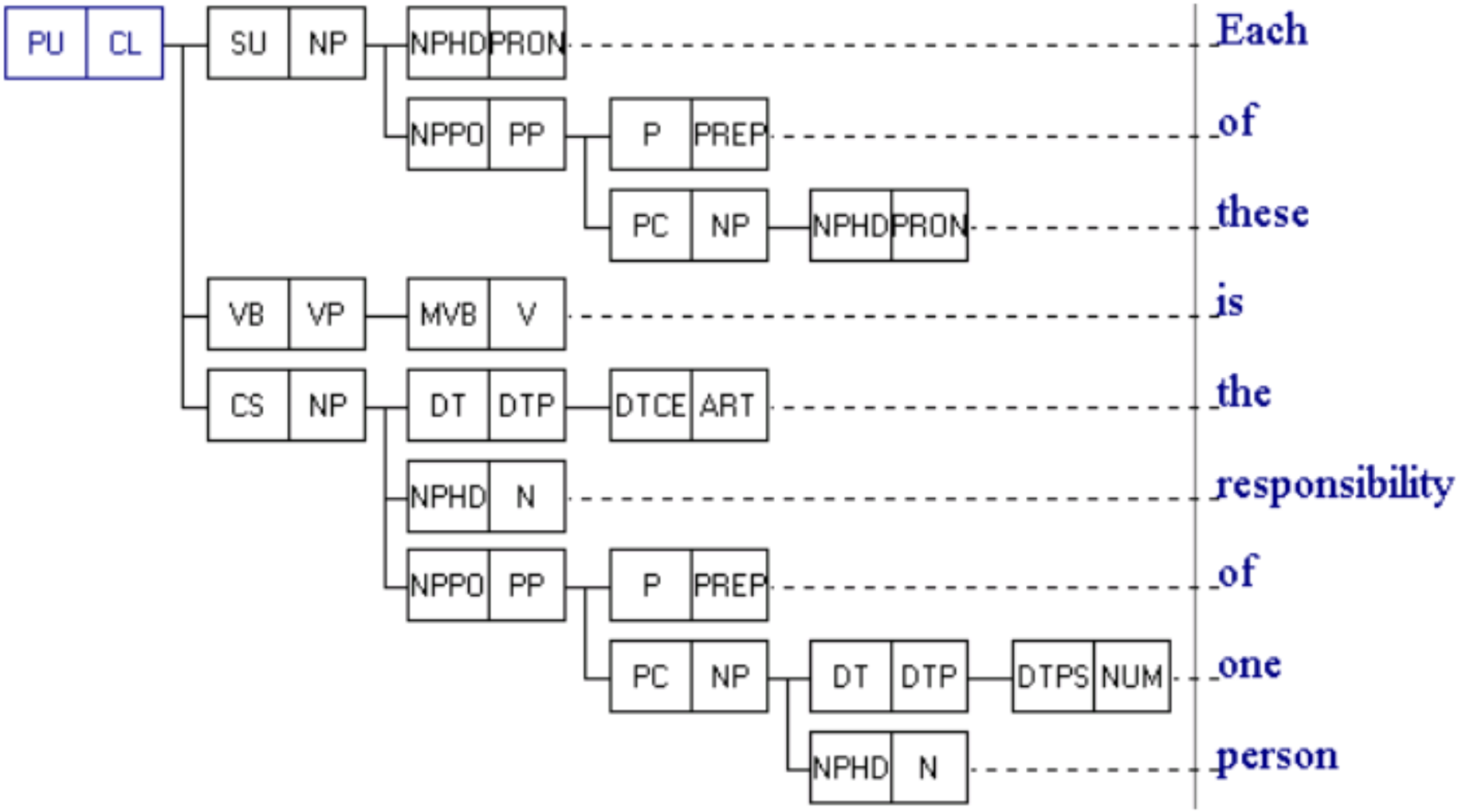
Spoken (300)	**Dialogues** (180)	**Private** (100)	Conversations (90) Phonecalls (10)
		Public (80)	Class Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Cross-examinations (10) Business Transactions (10)
	Monologues (120)	**Unscripted** (70)	Commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)

Written (200)	**Non-printed** (50)	**Student Writing** (20)	Student Essays (10) Exam Scripts (10)
		Letters (30)	Social Letters (15) Business Letters (15)
	Printed (150)	**Academic** (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Popular (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Reportage (20)	Press reports (20)
		Instructional (20)	Administrative Writing (10) Skills/hobbies (10)
		Persuasive (10)	Editorials (10)
		Creative (20)	Novels (20)

ICE

- Textual Markup
  - In written texts, features of the original layout are marked, including sentence and paragraph boundaries, headings, deletions, and typographic features.
  - Spoken texts are transcribed orthographically, and are marked for pauses, overlapping strings, discourse phenomena such as false starts and hesitations, and speaker turns.
- Wordclass Tagging
  - ICE texts are automatically tagged for wordclass by the TOSCA Tagger

- Syntactic parsing



http://ice-corpora.net/ice/annotate.htm

- The following corpora are available free (under Licence) to download from this site:

- CANADA (ICE-CAN - 1m words, lexical)
  JAMAICA (ICE-JA - 1m words, lexical)
  HONG KONG (ICE-HK - 1m words, lexical)
  EAST AFRICA (ICE-EA - Kenya & Tanzania,)
  INDIA (ICE-IND - 1m-words, lexical)
  SINGAPORE (ICE-SIN - 1m words, lexical)
  PHILIPPINES (ICE-PHI - 1m words, lexical)
  USA (ICE-USA, written component - c.400,000 words, lexical)
  IRELAND (ICE-IRL - 1m words, lexical)
  SPICE-IRELAND (SPICE-IRL - c.600,000 words with prosodic and pragmatic annotation)

- The following corpora are also available,

- GREAT BRITAIN (ICE-GB - 1m words, POS-tagged and parsed, distributed with ICECUP retrieval software)
  NEW ZEALAND (ICE-NZ - 1m words, lexical)
  SRI LANKA (ICE-SL - written component; lexical and POS-tagged with CLAWS C7 tagset)
  NIGERIA (ICE-NG - written component).

Special purpose Corpus

- สร้างเพื่อเป็นตัวแทนของ sublanguage ที่ต้องการศึกษา
- วางวัตถุประสงค์ของการใช้ corpus นั้น : ทำประมวลศัพท์, ESP

ทำประมวลศัพท์

- corpus ต้องมี term ใน subfield นั้นครบถ้วน
- corpus ต้องให้ความกระจ่างในการหาความหมายของ term
- เลือกประเภทของ text type ให้เหมาะ
- วิเคราะห์จากผู้ส่งสารและผู้รับสาร
  - 1. ใช้และเข้าใจในหมู่ specialist
  - 2. ใช้เพื่อสอนหรือสื่อสารระหว่าง speiclaist กับผู้เริ่มต้นหรือนักเรียนใน field
  - 3. ใช้สื่อสารกับสาธารณะ ผู้เขียนอาจไม่ใช่ specialist ไม่ควรนำมา