Understanding corpus linguistics research: useful definitions

"A corpus is a collection of machine readable, authentic texts, which is sampled to be representative of a particular language or language variety." (McEnery et al., 2006: 5)

Understanding corpus linguistics research: What is corpus linguistics?

- It is an empirical (experimental) approach
  - An analysis of actual patterns of use in target texts
- It uses a corpus of natural texts as the basis for analysis
  - Corpus = a representative sample of target language stored as an electronic database (plural = "corpora")
- It relies on computer software for analysis
  - Results are generated using automatic and interactive techniques
- It depends on both quantitative and qualitative analytical techniques
  - Observations are counted and results are interpreted
Understanding corpus linguistics research: A brief history of corpora and corpus tools

~1960s
- Noam Chomsky puts grammar, computers, (and intuition) into the spotlight
- "Syntactic Structures" (1957)
- "Aspects of the Theory of Syntax" (1965)
- Intuition and experience dictate much of dictionary and textbook content
- Growth in computer technologies leads to the development of...
  - authentic language databases (corpora)
    - Brown Corpus (Henry Kucera and W. Nelson Francis, 1961)

Brown Corpus Sample
A01 0010 The Fulton County Grand Jury said Friday an investigation of Atlanta’s recent primary election produced "no evidence" that A01 0020 any irregularities took place. The jury further said in term-end A01 0030 of Atlanta’s recent primary election produced "no evidence" that A01 0040 any irregularities took place. The jury further said in term-end A01 0050 charge of the election, "deserves the praise and thanks of the A01 0060 City of Atlanta" for the manner in which the election was conducted. A01 0070 The September-October term jury had been charged by Fulton A01 0080 Superior Court Judge Durwood Pye to investigate reports of possible A01 0090 "irregularities" in the hard-fought primary which was won by A01 0100 Mayor-nominate Ivan Allen Jr. "Only a relative handful

Discon (Clark, 1966)

Discon. Purpose of program: Concordance-making. This program is simply the well-known DISCON, originally written for the 7090 and converted to the 7040 by R. L. Price of Roswell Park Memorial Institute, Buffalo, New York, then converted to the 7044 here by Marjorie Schultz. Type and format of output: Punched cards: 6 cols. 1D, the rest data. Programming language used: Fortran IV. Required hardware: IBM 7044. Running time: Approx. 4 min. per 1000 lines of poetry, exclusive of priming time.

Correspond with Roger Clark or Lewis Savin, 123 W Hellems, University of Colorado, Boulder, Colorado.

Understanding corpus linguistics research:
A brief history of corpora and corpus tools

~1970s-1980s
- creation of more corpora
  - The Lancaster/Oslo-Bergen Corpus (LOB) (1978)
  - Collins Birmingham University International Language Database (COBUILD) (1980)
- rapid growth in grammar/vocabulary studies based on corpora
  - Collins COBUILD English Language Dictionary (1987)
  - development of easier to use (2nd generation) software tools
    - e.g. MicroConcord (Scott & Johns, 1993)

1990s
- further growth in corpus-building
  - British National Corpus (BNC) (1995)
- development of tagging and annotation tools
  - CLAWS Tagger Ver. 4 (1994)
- development of (3rd generation) corpus-tools
  - WordSmith (1996-2013)
Understanding corpus linguistics research:
A brief history of corpora and corpus tools

~1990s
- creation of corpus-based dictionaries, textbooks, and grammar books
- Collins COBUILD Dictionary - 2nd Edition
- Longman Dictionary of Contemporary English - 3rd Edition
- (Macmillan English Dictionary for Advanced Learners - 2002)

2000s to today
- Continued development of easy-to-use 3rd-generation software tools
  - e.g. WordSmith Tools, AntConc
- Creation of server-based (4th-generation) software tools
  - e.g. COCA, Sketch Engine, CQPWeb
- Introduction of Data-Driven Learning (DDL) in the classroom
  - e.g. Tim Johns, Alex Boulton
- Creation of learner corpora
  - e.g. Ute Romer, Yukio Tono, Ruiying Yang
- Research on corpus-based vocabulary learning
  - e.g. Paul Nation

Corpus of Contemporary American English (COCA)
http://corpus.byu.edu/coca/
Understanding corpus linguistics research: From 1 m word corpora to 1 t word corpora

Number of Words

Understanding corpus linguistics research: From principled corpora to opportunistic corpora

Number of Words

Understanding corpus linguistics research: From corpora in files to corpora in ‘black boxes’

Understanding corpus linguistics research: insights from corpus data

Insights from corpus data: Writing conference abstracts

Total presentations: 99

<table>
<thead>
<tr>
<th></th>
<th>min (words)</th>
<th>max (words)</th>
<th>ave (words)</th>
<th>std</th>
<th>tokens</th>
<th>types</th>
</tr>
</thead>
<tbody>
<tr>
<td>titles</td>
<td>4</td>
<td>14</td>
<td>9</td>
<td>2.44</td>
<td>923</td>
<td>432</td>
</tr>
<tr>
<td>summaries</td>
<td>12</td>
<td>54</td>
<td>47</td>
<td>5.72</td>
<td>4698</td>
<td>1346</td>
</tr>
<tr>
<td>abstracts</td>
<td>78</td>
<td>261</td>
<td>207</td>
<td>45.25</td>
<td>20477</td>
<td>3146</td>
</tr>
</tbody>
</table>

Shortest?

Active reading with CALL

Longest?

Enhancing the CALL experience through Data-Driven-Learning: From words to texts and beyond
Insights from corpus data:
Writing conference abstracts

Total presentations: 99

Abstracts (frequency curve)

Abstracts (frequency bands)

Abstracts (frequency 'heat map')

Most frequent words

<table>
<thead>
<tr>
<th>rank</th>
<th>titles</th>
<th>summaries</th>
<th>abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>and</td>
<td>and</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>learning</td>
<td>the</td>
<td>and</td>
</tr>
<tr>
<td>3</td>
<td>the</td>
<td>of</td>
<td>of</td>
</tr>
<tr>
<td>4</td>
<td>of</td>
<td>to</td>
<td>to</td>
</tr>
<tr>
<td>5</td>
<td>for</td>
<td>in</td>
<td>in</td>
</tr>
<tr>
<td>6</td>
<td>to</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>7</td>
<td>in</td>
<td>will</td>
<td>students</td>
</tr>
<tr>
<td>8</td>
<td>with</td>
<td>students</td>
<td>for</td>
</tr>
<tr>
<td>9</td>
<td>a</td>
<td>this</td>
<td>will</td>
</tr>
<tr>
<td>10</td>
<td>english</td>
<td>for</td>
<td>learning</td>
</tr>
</tbody>
</table>
Insights from corpus data:
Writing conference abstracts

Highest ranked keywords

<table>
<thead>
<tr>
<th>rank</th>
<th>titles</th>
<th>summaries</th>
<th>abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>learning</td>
<td>students</td>
<td>students</td>
</tr>
<tr>
<td>2</td>
<td>efl</td>
<td>presentation</td>
<td>learning</td>
</tr>
<tr>
<td>3</td>
<td>digital</td>
<td>learning</td>
<td>language</td>
</tr>
<tr>
<td>4</td>
<td>online</td>
<td>will</td>
<td>english</td>
</tr>
<tr>
<td>5</td>
<td>active</td>
<td>efl</td>
<td>presentation</td>
</tr>
<tr>
<td>6</td>
<td>english</td>
<td>english</td>
<td>learners</td>
</tr>
<tr>
<td>7</td>
<td>moodle</td>
<td>learners</td>
<td>online</td>
</tr>
<tr>
<td>8</td>
<td>learners</td>
<td>online</td>
<td>will</td>
</tr>
<tr>
<td>9</td>
<td>using</td>
<td>language</td>
<td>presenter</td>
</tr>
<tr>
<td>10</td>
<td>language</td>
<td>presenter</td>
<td>teachers</td>
</tr>
</tbody>
</table>

Stat: LL; p< 0.05 + Bonferroni correction; ref. corpus = AmE06 (Potts and Baker, 2012)

Most frequent n-grams (n=2)

<table>
<thead>
<tr>
<th>rank</th>
<th>titles</th>
<th>summaries</th>
<th>abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>active learning</td>
<td>will be</td>
<td>of the</td>
</tr>
<tr>
<td>2</td>
<td>in a</td>
<td>this presentation</td>
<td>in the</td>
</tr>
<tr>
<td>3</td>
<td>learning in</td>
<td>the</td>
<td>will be</td>
</tr>
<tr>
<td>4</td>
<td>study abroad</td>
<td>presentation will</td>
<td>this presentation</td>
</tr>
<tr>
<td>5</td>
<td>in a</td>
<td>the</td>
<td>students</td>
</tr>
<tr>
<td>6</td>
<td>blended learning</td>
<td>of the</td>
<td>students to</td>
</tr>
<tr>
<td>7</td>
<td>extensive reading</td>
<td>will demonstrate</td>
<td>on the</td>
</tr>
<tr>
<td>8</td>
<td>language learning</td>
<td>the presenter</td>
<td>can be</td>
</tr>
<tr>
<td>9</td>
<td>effectiveness of</td>
<td>of a</td>
<td>the presenter</td>
</tr>
<tr>
<td>10</td>
<td>an online</td>
<td>can be</td>
<td>language learning</td>
</tr>
</tbody>
</table>

Stat: LL; p< 0.05 + Bonferroni correction; ref. corpus = AmE06 (Potts and Baker, 2012)

Insights from corpus data:
Writing conference abstracts

Highest ranked key n-grams (n=2)

<table>
<thead>
<tr>
<th>rank</th>
<th>titles</th>
<th>summaries</th>
<th>abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>active learning</td>
<td>this presentation</td>
<td>this presentation</td>
</tr>
<tr>
<td>2</td>
<td>extensive reading</td>
<td>presentation will</td>
<td>the presenter</td>
</tr>
<tr>
<td>3</td>
<td>language learning</td>
<td>the presenter</td>
<td>language learning</td>
</tr>
<tr>
<td>4</td>
<td>blended learning</td>
<td>will demonstrate</td>
<td>the students</td>
</tr>
<tr>
<td>5</td>
<td>study abroad</td>
<td>will be</td>
<td>students to</td>
</tr>
<tr>
<td>6</td>
<td>learning in</td>
<td>active learning</td>
<td>presentation will</td>
</tr>
<tr>
<td>7</td>
<td>effectiveness of</td>
<td>presenter will</td>
<td>the classroom</td>
</tr>
<tr>
<td>8</td>
<td>google docs</td>
<td>japanese university</td>
<td>english language</td>
</tr>
<tr>
<td>9</td>
<td>efl learners</td>
<td>will introduce</td>
<td>active learning</td>
</tr>
<tr>
<td>10</td>
<td>learning together</td>
<td>language learning</td>
<td>in japan</td>
</tr>
</tbody>
</table>

Stat: LL; p< 0.05 + Bonferroni correction; ref. corpus = AmE06 (Potts and Baker, 2012)

Insights from corpus data:
Writing conference abstracts

Graph of abstracts -> top 20 key-n-grams (n=2)

Doing corpus linguistics:
applications in morphology (vocabulary analysis)
Applications in morphology:
What words do you need to know to read any text in English?

Applications in morphology:
What words do you need to know to read any text in English?

Applications in morphology:
What words do you need to know to read any text in English?

Applications in morphology:
What words do you need to know to read any text in English?
Applications in morphology:
What words do you need to know to read any text in English?

"It is important that learners have access to lists of high-frequency and academic words and are able to obtain frequency information from dictionaries." (p. 219)

"Priority should be given to high-frequency words and to words that clearly fulfill language use needs. (p. 303)

Applications in syntax:
Do scientists always write in the passive voice?

"We show that ..."  "It was shown that ..."

Doing corpus linguistics:
applications in syntax (second-language learning)

Applications in syntax:
What does 'blood' mean in Shakespeare's plays?

Doing corpus linguistics:
applications in semantics (literature studies)

<table>
<thead>
<tr>
<th>Active Form</th>
<th>Freq</th>
<th>Location</th>
<th>Passive Form</th>
<th>Freq</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>we found</td>
<td>29</td>
<td>Dis</td>
<td>was/were found</td>
<td>26</td>
<td>Res/Dis</td>
</tr>
<tr>
<td>we show</td>
<td>20</td>
<td>Dis</td>
<td>is/are shown</td>
<td>26</td>
<td>Met/Res</td>
</tr>
<tr>
<td>we have</td>
<td>18</td>
<td>Dis</td>
<td>is/are had</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>we thank</td>
<td>10</td>
<td>Ack</td>
<td>is/are thanked</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>we observed</td>
<td>15</td>
<td>Dis</td>
<td>was/were observed</td>
<td>45</td>
<td>Res/Dis</td>
</tr>
<tr>
<td>we used</td>
<td>14</td>
<td>Intro/Meth/Res/Dis</td>
<td>was/were used</td>
<td>78</td>
<td>Met</td>
</tr>
<tr>
<td>we investigated</td>
<td>11</td>
<td>Abs/Int/Res/Dis</td>
<td>was/were investigated</td>
<td>6</td>
<td>Abs/Int/Res/Dis</td>
</tr>
</tbody>
</table>

Dermatology Corpus (52 texts, 89246 tokens)
Applications in syntax:
What does 'blood' mean in Shakespeare's plays?

<table>
<thead>
<tr>
<th>Example quotes</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Her blood is settled, and her joints are stiff.</td>
<td>Romeo &amp; Juliet</td>
</tr>
<tr>
<td>My blood for your rude brawls doth lie a-bleeding.</td>
<td>Romeo &amp; Juliet</td>
</tr>
<tr>
<td>To kiss these lips, I will appear in blood.</td>
<td>Antony &amp; Cleopatra</td>
</tr>
<tr>
<td>When the blood burns, how prodigal the soul.</td>
<td>Hamlet</td>
</tr>
<tr>
<td>Make thick my blood.</td>
<td>Macbeth</td>
</tr>
<tr>
<td>And on thy blade and dudgeon gouts of blood.</td>
<td>Macbeth</td>
</tr>
<tr>
<td>There's blood upon thy face.</td>
<td>Macbeth</td>
</tr>
<tr>
<td>I do confess the vices of my blood.</td>
<td>Othello</td>
</tr>
</tbody>
</table>

Shakespeare Tragedies (5 plays, 109,999 tokens)

Doing corpus linguistics:
applications in pragmatics (conversational analysis)

Applications in pragmatics:
How do you say what you "want" to an operator?

- How do people make requests during a telephone call?
- Who asks most of the questions?
- How do you interrupt someone?
- How do you finish a conversation?
- ...

Examples of politeness (customer)  Examples of politeness (operator)

ok I want to leave on the ...  ... did you want me to verify it ...
... basically what I want to know is ...  ... did you want to leave Chicago ...
I don’t know if I want to ...  ... how soon do you want to ...
I have a couple of things I want to ask ...  ... do you want to ...
... it’s not giving me the fare I want.  ... if you want to go ...

Travel Operator Corpus (50 interactions, 43050 tokens)

Applications in language variation:
Do all scientists talk like the guys in The Big Bang Theory?

Doing corpus linguistics:
applications in language variation
(discourse analysis)
Do all scientists talk like the guys in *The Big Bang Theory*?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Big Bang Theory</th>
<th>Physics</th>
<th>Biology</th>
<th>Chemistry</th>
<th>Math</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>you</td>
<td>field</td>
<td>is</td>
<td>is</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>i</td>
<td>is</td>
<td>cells</td>
<td>energy</td>
<td>zero</td>
</tr>
<tr>
<td>3</td>
<td>oh</td>
<td>zero</td>
<td>cell</td>
<td>electron</td>
<td>equation</td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>magnetic</td>
<td>dna</td>
<td>molecule</td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>sheldon</td>
<td>current</td>
<td>gene</td>
<td>electrons</td>
<td>solution</td>
</tr>
<tr>
<td>6</td>
<td>re</td>
<td>charge</td>
<td>ft</td>
<td>here</td>
<td>y</td>
</tr>
<tr>
<td>7</td>
<td>okay</td>
<td>electric</td>
<td>fre</td>
<td>orbital</td>
<td>function</td>
</tr>
<tr>
<td>8</td>
<td>t</td>
<td>this</td>
<td>here</td>
<td>this</td>
<td>this</td>
</tr>
<tr>
<td>9</td>
<td>yeah</td>
<td>direction</td>
<td>protein</td>
<td>molecules</td>
<td>omega</td>
</tr>
<tr>
<td>10</td>
<td>leonard</td>
<td>here</td>
<td>genome</td>
<td>orbitals</td>
<td>matrix</td>
</tr>
</tbody>
</table>

Highest Ranked Keywords in Discipline-Specific Spoken Corpora
Stat: LL; p < 0.05 + Bonferroni correction; Ref. corpus = AmE06 (Potts and Baker, 2012)

Traveling to the future in corpus linguistics:
A growing interest in "big data"

**The Rise of Big Data**
How It’s Changing the Way We Think About the World

The value of big data to the UK economy

[Image of a person and chart showing growth]

Traveling to the future in corpus linguistics:
A growing interest in "digital" humanities

Growth of publications on or citing the topic Digital Humanities or Humanities computing (Scharnhorst, 2014)

2014 over 140

1968 less than 5

Traveling to the future in corpus linguistics:
A growing interest in (social) data analytics

Tableau
https://www.tableau.com

Traveling to the future in corpus linguistics:
A growing interest in (social) data analytics

DataWrapper
https://www.datawrapper.de
Traveling to the future in corpus linguistics: A growing interest in (social) data analytics

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.

Traveling to the future in corpus linguistics: The Scientific Research Process

The “Corpus Approach”

1. Choose a topic to investigate
2. Choose a target area
3. Find a ready-built corpus in the target area
4. Design your own custom corpus
5. Conduct an empirical investigation
6. Decide a sampling procedure
7. Collect, clean, tag, and/or annotate the data
8. Identify a gap
9. Collect/Analyze data
10. Discuss the results
11. Design an experiment
12. Review the literature

Microsoft Corp., Automatic translation for Japanese

Microsoft announced on July 7 that the service for automatically translating speech in different languages in real time was made available in Japanese. Using free smartphone (smartphone) applications, you can interact while listening to the translated voice, and watch the translated subtitles in handy smartphone. It corresponds to English, Chinese, Spanish etc. We applied artificial intelligence (AI) related technology and increased accuracy.
Traveling to the future in corpus linguistics:
The Scientific Research Process

- Step 4: Design an experiment
  - Materials (corpora)
    - BNC, MICASE, MICUSP, BROWN, LOB, ...
    - scanned/OCR textbooks, PDF research papers, ...
  - Materials (software tools)
    - AntConc, WordSmith Tools, COCA, BNC Web, ...
    - TagAnt, Stanford POS Tagger, Stanford Parser, ...
  - Methods
    - concordancing, concordance plotting
    - clusters/n-gram analysis
    - collocation analysis
    - word/keyword analysis
    - statistical analysis/interpretation of the results

- Step 5: Collect/analyze the data
- Step 6: Discuss the results

Eye-Tracking Research
Detecting eye-fixations with EyeChat

**NS1 - fixated words (based on EyeChat)**
- 'Going overseas for university study is an exciting prospect for many people. But while it may offer some advantages, it is probably better to stay home because of the difficulties a student inevitably encounters living and studying in a different culture.'

To what extent do you agree or disagree with this statement? Give reasons for your answer and include any relevant examples from your knowledge or experience.

Discuss with your partner about this topic.

Interactive text-enhancement with EyeChat

Text fixations highlighted and ‘enhanced’

Chat fixations highlighted and ‘enhanced’

EyeChat
Traveling to the future in corpus linguistics: Data-Mining and Network Analysis

FireAnt: Filter, Identify, Report, and Export Analysis Tool

Twitter collector in action

FireAnt visualization output (exported to Gephi)

86

Analysis of research article titles in English for Specific Purposes

Data-Mining and Network Analysis: FireAnt: Filter, Identify, Report, and Export Analysis Tool

Politics as seen through social media

1st US Presidential Debate 2016

Ted Cruz @tedcruz

We need to unite to defend freedom and restore the Constitution #Debates2016
12:18 PM - 27 Sep 2016
1,191 Retweets 2,762 Likes

Alana Massey @alana

Tonight, one of history’s most accomplished politicians must share a stage with a man who hinted that someone should shoot for 4 weeks ago
7:00 AM - 27 Sep 2015
8 Retweets 1 Favorite

The anatomy of a tweet
Data-Mining and Network Analysis:
Politics as seen through social media

Tweet histories of Clinton and Trump
November 2015-May 2016

Clinton tweets: 3211
Trump tweets: 3205

Trump Tweet history (June 2016-June 2017)
Total tweets: 3016 (no retweets)

Clinton Positive-Negative Network (1000 tweets 2016/05/18)
Data Analysis: FireAnt; User mentions > 0 (628 'friends')
Rendering: Gephi 0.9.1; Force Atlas Layout (Default settings + 10000 repulsion)

Trump Positive-Negative Network (1000 tweets 2016/05/18)
Data Analysis: FireAnt; User mentions > 0 (360 'friends')
Rendering: Gephi 0.9.1; Force Atlas Layout (Default settings + 10000 repulsion)

Summary and Questions:
What next?

Understanding corpus linguistics
- What is corpus linguistics?
- How did it advance between the 1960s and today?
- What can it tell us about language and people and society?

Doing corpus linguistics
- How can we use corpus linguistics to inform us about...
  - morphology, syntax, semantics, pragmatics, language variation
  - vocabulary analysis, second language learning, literature studies, conversational analysis, and discourse analysis

Traveling to the future in corpus linguistics
- What new areas of corpus linguistics research are emerging?
- What corpus data sources are being used?
- What new corpus tools are being developed to help us?