Corpus Wars:
Struggling to Find New Ways of
Analyzing and Teaching Vocabulary

Laurence Anthony
Center for English Language Education in Science and Engineering
Faculty of Science and Engineering, Waseda University
http://www.antlab.sci.waseda.ac.jp/
anthony@antlab.sci.waseda.ac.jp

ICTATLL Workshop 2007: ICT in Analysis, Teaching, and Learning of Languages
Aug. 1st, 2007

Outline

Corpus Linguistics: A New Hope?
- challenges in the study of grammar and vocabulary
  - grammar vs. vocabulary
  - lexical phrases
- corpus-based vs. corpus-driven approaches
- Vocabulary strikes back?
  - words vs. lexical phrases
  - contextualized vs. de-contextualized learning
  - simplified texts vs. authentic texts

Outline

Return of Corpus Linguistics: ICT for the future
- New tools for corpus linguistics
  - POS taggers, parsers, and concordance software
- New tools for vocabulary analysis
  - tools to identify vocabulary difficulty and text readability
  - authoring tools for extensive reading
  - tools for (automated) simplification of texts

Corpus Linguistics:
A New Hope?

~1960s~
- Noam Chomsky puts grammar in the spotlight
  - “Syntactic Structures” (1957)
  - “Aspects of the Theory of Syntax” (1965)
- contents of dictionaries, textbooks, etc. decided largely on intuition and experience but not heavily on statistics
- growth in computer technologies leads to development of language databases (corpora) but almost no software to analyze them
  - e.g. Brown Corpus (1961)
  - Henry Kucera and W. Nelson Francis

Corpus Linguistics:
A New Hope?

Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so widely skewed that the description would be no more than a mere list.

Noam Chomsky
Third Texas Conference on Problems of Linguistic Analysis in English, 1958, p. 159.
Austin: University of Texas (1962).

Corpus Linguistics:
A New Hope?

A01 0010 The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta’s recent primary election produced “no evidence” that
A01 0030 any irregularities took place. The jury further said in term-end
A01 0040 that an investigation reports of possible
A01 0050 “irregularities” in the hard-fought primary which was won by
A01 0060 Mayor-nominate Ivan Allen Jr. “Only a relative handful

Brown Corpus Sample
Corpus Linguistics: A New Hope?

~1970s~
- "The OSTI Report" (1970)
  - Final Report to the Office for Scientific and Technical Information (OSTI) on the Lexis Research Project
  - John Sinclair, Susan Jones, and Robert Daley
- A span of 4(5) words is useful to identify collocations
- ATOL programming and CLOC software tools for text analysis and collocation identification
- All forms of lemmas do NOT show the same collocational patterns (semantic value)
- Words are NOT the basic building blocks of language - instead we should consider lexical units.

~1970s-1980s~
- Rapid growth in language studies based on corpora
  - The Lancaster/Oslo-Bergen Corpus (LOB) (1978)
  - Collins Birmingham University International Language Database (COBUILD) (1980)
  - Collins COBUILD English Language Dictionary (1987)
- Debate on corpus driven vs. corpus based approaches

~1990s-2000s~
- Rapid growth in dictionaries, textbooks, grammar books heavily based on results from corpora
  - British National Corpus (BNC) (1995)
- Debate on corpus driven vs. corpus based approaches

Date | Studies
--- | ---
To 1965 | 10
1966-1970 | 20
1971-1975 | 30
1976-1980 | 80
1981-1985 | 160
1985-1991 | 320

Joanassson, 1991

Corpus Linguistics: A New Hope?

Corpus driven vs. corpus based approaches
- The Birmingham School (corpus-driven)
  - Lexical items are the basic unit of meaning
    - Words, collocations, idioms
    - By the way, long time no see,
    - As far as I ___, no only ___ but ___
- Collocational principles (Sinclair, 1991)
  - Open-choice principle
    - Slot and fill according to grammar categories
  - Idiom principle
    - "The large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments."

Elena Tognini-Bonelli
Corpus Linguistics at Work (2001)

"While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach."

Wolfgang Teubert

"All the material included in a corpus, whether spoken, written or gathered along any intermediate dimension is assumed to be taken from genuine communications of people going about their normal business."

Elena Tognini-Bonelli
Corpus Linguistics at Work (2001)
Corpus Linguistics: A New Hope?

- corpus driven vs. corpus based approaches
  - The Birmingham School (corpus-driven)
    - "Pattern Grammar: A corpus-driven approach to the lexical grammar of English"
      - Susan Hunston and Gill Francis (2000)
    - patterns in grammar relate to meanings in expressions
      - words and grammar are mutually dependent
        - e.g. Only nine "VERB + so" patterns
          - assume so, believe so, fear so, hope so imagine so, presume so, say so, suspect, think so
  - The Lancaster School (corpus-based)
    - "The corpus based approach ... draws upon authentic or real texts..."
    - "Corpora are used mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study."
    - "Additional advantages of the corpus-based approach are that a corpus can find differences that intuition alone cannot perceive."

Corpus-Based Language Studies (p. 6)

Additional advantages of the corpus-based approach are that a corpus can find differences that intuition alone cannot perceive.

- Corpus-based approaches rely on corpora that are:
  - balanced and representative (e.g. BNC)
  - not necessary for corpus-driven studies
  - small or large (e.g. LOB or BNC)
    - large corpora are necessary for corpus-driven studies
  - usually annotated (e.g. LOB or BNC)
    - not necessary (should be avoided?) in corpus driven studies
  - usually annotated (e.g. LOB or BNC)
    - not necessary (should be avoided?) in corpus driven studies

Corpus-Based Language Studies (p. 10-11)

Vocabulary strikes back?

- Nation's Balanced Language Course
  - meaning-focused input (in reading and writing)
    - "...learning from meaning-focused input can best occur if learners are familiar with at least 95% of the running words in the input they are focusing on."
  - language-focused input
    - "language learning benefits if there is an appropriate amount of carefully focused deliberate teaching and learning of language items"
  - meaning-focused output
  - fluency development

Nation, Vocabulary (2000)
Vocabulary strikes back?

- Nation’s *Word Families*
  - “A word family consists of a headword, its inflected forms, and its closely related derived forms.”
  - battle → battle, battled, battles, battling, battler, battlers
  - “The high frequency words of a language are clearly so important that considerable time should be spent on them by teachers and learners”
    - top 1000 words → 74% of running words
    - top 2000 words → 78% of running words
    - top 1000 + acad. words → 85% of running words

- What does Nation mean when he says “*know a word*”?
  - **Form**
    - What does the word sound like?
    - How is the word pronounced?
    - What does the word look like?
    - How is the word written or spelled?
    - What parts are recognisable in this word?
    - What word parts are needed to express the meaning?

- What does Nation mean when he says “*know a word*”?
  - **Meaning**
    - What meaning does this word form signal?
    - What word form can be used to express this meaning?
    - What is included in the concept?
    - What items can the concept refer to?
    - What other words does this make us think of?
    - What other words could we use instead of this one?

- What should we do?
  - spend time on high-frequency words
    - word cards
    - bilingual dictionaries
    - simplified texts
  - don’t waste time studying low-frequency words
    - develop lower frequency words through exposure to the language (via an extended reading program) and strategies

- What does Nation mean when he says “*know a word*”?
  - **Use**
    - In what patterns does (or must) the word occur?
    - What words or types of words occur (or must occur) with this one?
    - Where, when, and how often would we expect to meet this word?
    - Where, when, and how often can we use this word?

- What does Nation mean when he says “*know a word*”?
  - ...
“It is worth noting that there are principles that some teachers and course designers follow that go against research findings. These include: ‘All vocabulary learning should occur in context,’ ‘The first language should not be used as a means of presenting the meaning of a word,’ ‘Vocabulary should be presented in lexical sets,’ ‘Monolingual dictionaries are preferable to bilingual dictionaries,’ ‘Most attention should be paid to the first presentation of a word,’ and ‘Vocabulary learning does not benefit from being planned, but can be determined by the occurrence of words in texts, tasks and themes.’”

I.S.P Nation
Vocabulary (2005, p. 384)

What should we do?
- spend time on high-frequency words
  - word cards, morphology, translation, repetition
- don’t waste time studying low-frequency words
  - develop lower frequency words through exposure to the language (via extended reading) and strategies

Hardware and software tools to work with vocabulary
- Paul Nation’s Range
- Tom Cobb’s VocabProfile
- Shinichi Shimizu’s JACET 8000 Level Marker
**Vocabulary strikes back?**

- Nation’s view of concordancers and collocations
  - “Learners need **training** in how to use concordancers, and the data obtained from concordancers needs to be **comprehensible** to the learner.”
  - “**Frequent collocations** deserve attention in the classroom if their frequency is equal to or higher than other high frequency words.”

**Return of Corpus Linguistics:**

- **ICT for the future?**
  - Increasing use of **annotated** corpora
    - We need **easier-to-use** and **more accurate** taggers and parsers
  - Increasing use of corpora in language learning
    - We need **easier-to-use** and **more sophisticated** concordance software
  - Increasing importance of vocabulary and extended reading programs
    - We need **better tools** to analyze vocabulary
      - to calculate text readability
      - to identify good exemplars of word usage
      - to **author** extended readers and **simplify** existing texts

**Corpus Linguistics:**

- **A New Hope?**
  - **AntCLAWS** (A multiplatform CLAWS Interface)
    - http://www.antlab.sci.waseda.ac.jp
  - **AntConc** (A tag aware concordancer)
    - http://www.antlab.sci.waseda.ac.jp
  - **AntVocabCheck** (A vocabulary analyzer)
    - http://www.antlab.sci.waseda.ac.jp
  - **AntDictionaryEntry** (A automated exemplar creator)
    - http://www.antlab.sci.waseda.ac.jp
Return of Corpus Linguistics: ICT for the future?

Conclusion

- Corpus linguistics is a new hope in language analysis and learning
- But...
  - the tools for carrying out corpus studies are often limited, difficult to use, and out-dated
  - novel tools for advanced analysis are necessary
- And we shouldn’t forget language learning...
  - vocabulary (in addition to lexis!) is an essential part of language learning
  - more tools are needed for studying vocabulary, and authoring/simplifying texts for language learning