How to Make a Dictionary:
Computational Lexicography

Dafydd Gibbon

B.A. British And American Studies
Basic Module 2

Winter Semester 2007/2008
Overview

• REVIEW OF LEXICOGRAPHY:
  – LEXICAL DATA ACQUISITION

• CONCORDANCES
  – SIMPLE KWIC CONCORDANCE

• COMPUTING A KWIC CONCORDANCE
  – From Text to KWIC Concordance

• DICTIONARY MAKING
  – KWIC: Dictionary Making
REVIEW OF LEXICOGRAPHY

PRINCIPLES
Criteria for Good Lexicography

• Quantity:
  – Completeness of coverage:
    • extensional coverage: number of entries
    • intensional coverage: number of types of lexical information

• Quality:
  – Correctness of information:
    • Types of lexical information
  – Consistency of structure:
    • Macrostructure
    • Microstructure
    • Mesostructure
Lexicographic workflow cycle
Lexicographic workflow cycle

Data acquisition:
• Recordings
• Text collection
• Concordance
• Dictionaries
• ...
Lexicographic workflow cycle

Data acquisition:
- Recordings
- Text collection
- Concordance
- Dictionaries
- ...

Lexicon construction:
- Metadata
- Information retrieval
- Linguistic analyses
Data acquisition:
• Recordings
• Text collection
• Concordance
• Dictionaries
• ...

Lexicon construction:
• Metadata
• Information retrieval
• Linguistic analyses

Access to data:
• Traditional print media
• Hyperlexicon: CD, internet
• Software with lexicon component:
  • word processing
  • speech processing
**Lexicographic workflow cycle**

**Data acquisition:**
- Recordings
- Text collection
- Concordance
- Dictionaries
- ...

**Lexicon construction:**
- Metadata
- Information retrieval
- Linguistic analyses

**Lexical evaluation:**
- Internal:
  - consistency
  - completeness
- External:
  - utility for the users

**Access to data:**
- Traditional print media
- Hyperlexicon: CD, internet
- Software with lexicon component:
  - word processing
  - speech processing
LEXICAL DATA ACQUISITION
From corpus to lexicon
From corpus to lexicon

Layer 1: Primary data (audio / video recording)
From corpus to lexicon

CORPUS

Layer 1: Primary data
(audio / video recording)

Layer 2: Secondary data
(transcription, annotation, metadata)
From corpus to lexicon

Layer 1: Primary data
(audio / video recording)

Layer 2: Secondary data
(transcription, annotation, metadata)

LEXICON

CORPUS

Lexicographic complexity
From corpus to lexicon

Layer 1: Primary data
(audio / video recording)

Layer 2: Secondary data
(transcription, annotation, metadata)

Layer 1: Corpus lexicon
(wordlist, concordance, HMM, ...)

Lexicographic complexity
From corpus to lexicon

**Layer 1: Primary data**
(audio / video recording)

**Layer 1: Corpus lexicon**
(wordlist, concordance, HMM, ...)

**Layer 2: Secondary data**
(transcription, annotation, metadata)

**Layer 2: Lexicon matrix**
(entries x data categories, no generalisations)

LEXICON

CORPUS

Lexicographic complexity
From corpus to lexicon

Layer 1: Corpus lexicon
(wordlist, concordance, HMM, ...)

Layer 2: Lexicon matrix
(entries x data categories, no generalisations)

Layer 3: Lexicon with selected generalisations
(procedurally optimised: semasiological, onomasiological)

Layer 1: Primary data
(audio / video recording)

Layer 2: Secondary data
(transcription, annotation, metadata)

Lexicographic complexity
From corpus to lexicon

Layer 1: Primary data
(audio / video recording)

Layer 2: Secondary data
(transcription, annotation, metadata)

Layer 3: Lexicon with selected generalisations
(procedurally optimised: semasiological, onomasiological)

Layer 4: Lexicon with generalisation hierarchies
(general, type, default inheritance)

Layer 1: Corpus lexicon
(wordlist, concordance, HMM, ...)

Layer 2: Lexicon matrix
(entries x data categories, no generalisations)
FROM CORPUS TO LEXICON.
CONCORDANCES
Concordance

• A KWIC (KeyWord In Context) concordance is a special kind of preliminary, corpus-based dictionary:
  – each word in a text corpus is paired with its contexts of occurrence in this corpus

• Note: Google is a special form of KWIC concordance:
  – Google as a concordance
Concordance

• A KWIC (KeyWord In Context) concordance is a special kind of preliminary, corpus-based dictionary:
  – each word in a text corpus is paired with its contexts of occurrence in this corpus

• Example text:

“My first sight of England was on a foggy March night in 1973 when I arrived on the midnight ferry from Calais.”

Bill Bryson: Notes from a Small Island
Alphabetically ordered KWIC

Keywords with right-hand contexts:

<table>
<thead>
<tr>
<th>1973</th>
<th>when i arrived</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>foggy march night</td>
</tr>
<tr>
<td>arrived</td>
<td>on the midnight</td>
</tr>
<tr>
<td>calais</td>
<td>was on a</td>
</tr>
<tr>
<td>england</td>
<td>from calais</td>
</tr>
<tr>
<td>ferry</td>
<td>first</td>
</tr>
<tr>
<td>first sight of england</td>
<td>march night in</td>
</tr>
<tr>
<td>from</td>
<td>calais</td>
</tr>
<tr>
<td>i</td>
<td>arrived on the</td>
</tr>
<tr>
<td>in</td>
<td>1973 when i</td>
</tr>
<tr>
<td>march</td>
<td>night in 1973</td>
</tr>
<tr>
<td>midnight</td>
<td>ferry from calais</td>
</tr>
<tr>
<td>my</td>
<td>first sight of</td>
</tr>
</tbody>
</table>

| night | in 1973 when |
| of | england was on |
| on | a foggy march |
| on | the midnight ferry |
| sight | of england was |
| the | midnight ferry from |
| was | on a foggy |
| when | i arrived on |
Concordancing on the web

• The first:
  – HyprLex
  – VerbMobil HyprLex

• Some more:
  – General information on concordancing
  – Corpus Linguistics
A KWIC CONCORDANCE ENGINE
KWIC concordance construction

1. CORPUS CREATION → 2. TOKENISATION →

3. KEYWORDLIST EXTRACTION  4. CONTEXT COLLATION →

5. KEYWORD SEARCH

6. OUTPUT FORMATTING
Simplest KWIC procedure

1. Corpus creation: make a corpus of texts in electronic format
2. Tokenisation (re-process each text):
   1. process punctuation marks
   2. break the text into context units (lines/sentences)
3. Keyword list extraction (all words in text)
4. Context collation (for each keyword)
5. Search for KWIC in corpus
6. Store output and format
   – for printing, hypertext (CD, web)
QUIZ

• What is a KWIC concordance?
• Which are the two main components of lexicon construction based on empirical data?
• Which layers of abstraction are involved in corpus acquisition?
• Which layers of abstraction are involved in lexicon construction? Describe them.
• Which layer do standard dictionary types typically belong to?
SIMPLE KWIC CONCORDANCE
KWIC: 1. Corpus collation

• My first sight of England was on a foggy March night in 1973 when I arrived on the midnight ferry from Calais.
KWIC procedure: 2. Tokenisation

• In the text:

*My first sight of England was on a foggy March night in 1973 when I arrived on the midnight ferry from Calais.*

• Process
  – upper case (capital) letters
  – punctuation marks

• To produce:

my first sight of england was on a foggy march night in 1973 when i arrived on the midnight ferry from calais
KWIC procedure: 3. Keyword List

- Replace each SP (space) sequence by a LF (linefeed) / NL (newline)
- Sort the list alphabetically
- Remove duplicate words
### KWIC procedure: 3. Keyword List

<table>
<thead>
<tr>
<th>my</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>1973</td>
</tr>
<tr>
<td>sight</td>
<td>when</td>
</tr>
<tr>
<td>of</td>
<td>i</td>
</tr>
<tr>
<td>england</td>
<td>arrived</td>
</tr>
<tr>
<td>was</td>
<td>on</td>
</tr>
<tr>
<td>on</td>
<td>the</td>
</tr>
<tr>
<td>a</td>
<td>midnight</td>
</tr>
<tr>
<td>foggy</td>
<td>ferry</td>
</tr>
<tr>
<td>march</td>
<td>from</td>
</tr>
<tr>
<td>night</td>
<td>calais</td>
</tr>
</tbody>
</table>
**KWIC procedure: 3. Keyword List**

| my first sight of england was on a foggy march night | in 1973 when i arrived on the midnight foggy ferry from i in | 1973 a arrived calais england ferry first foggy from i in | march midnight my night of on on on sight the was when |
KWIC procedure: 3. Keyword List

| my first sight of england was on a foggy march night | in 1973 when i arrived on the midnight ferry from calais | 1973 a arrived calais england ferry first foggy from i in March midnight my night of on on sight the was when | 1973 a arrived calais england ferry first foggy from i in March midnight my night of on on sight the was when | in March midnight my night of on on sight the was when |
KWIC procedure: 4. Contexts

• Pick context unit
  – left and right contexts
  – $m$ words at beginning
  – $n$ words at end

• Add $m$ boundary marks at beginning and and $n$ at end

• Split into units of length $m+1+n$
KWIC procedure: 4. Contexts

• Pick context unit
  – left and right contexts
  – $m$ words at beginning
  – $n$ words at end

• Add $m$ boundary marks at beginning and and $n$ at end

• Split into units of length $m+1+n$

  # my first
  my first sight
  first sight of
  sight of england
  england was on
  was on a
  on a foggy
  a foggy march
  foggy march night
  march night in
KWIC procedure: 4. Contexts

• Pick context unit
  – left and right contexts
  – $m$ words at beginning
  – $n$ words at end
• Add $m$ boundary marks at beginning and and $n$ at end
• Split into units of length $m+1+n$

# my first
my first sight
first sight of
sight of england
of england was
england was on
was on a
on a foggy
a foggy march
foggy march night
march night in
night in 1973
in 1973 when
1973 when i
when i arrived
i arrived on
arrived on the
on the midnight
the midnight ferry
midnight ferry from
ferry from calais
from calais #
KWIC procedure: 5. Search

• For example:
  – *on* is found in the middle of the following context units:
    • *was on a*
    • *arrived on the*
  – *arrived* is found in the middle of the following context units:
    • *i arrived on*
  – etc.
## KWIC procedure: 6. Output format

| 1973:     | in 1973 when |
| a:       | on a foggy  |
| arrived: | i arrived on |
| calais:  | from calais # |
| england: | of england was |
| ferry:   | midnight ferry from |
| first:   | my first sight |
| foggy:   | a foggy march |
| from:    | ferry from calais |
| i:       | when i arrived |
| in:      | night in 1973 |
| march:   | foggy march night |
| midnight:| the midnight ferry |
| my:      | # my first |
| night:   | march night in |
| of:      | sight of england |
| on:      | arrived on the |
| on:      | was on a |
| sight:   | first sight of |
| the:     | on the midnight |
| was:     | england was on |
| when:    | 1973 when i |
QUIZ

• What are the 6 main steps in KWIC concordance construction?
• Explain each of these steps:
  1.
  2.
  3.
  4.
  5.
  6.
COMPUTING A KWIC
CONCORDANCE
From Text to KWIC Concordance

TEXT CORPUS
KWIC procedure: 1. Preprocess

```perl
$wordlist = "" ;

while(<>) {
    chomp;
    s/e\./EG/ ;
    s/M\./MA/ ;
    tr/[.\;,\:""–)(]/ / ;
    tr/[A-Z]/[a-z]/ ;
    tr/\t/ / ;
    s/ */ /g ;
    $wordlist = $wordlist . $_;
}
```

NORMALISED TEXT
KWIC procedure: 2. Contexts

```perl
$contextlength = 5 ; @contextlist = () ; for ($i=0; $i<(@wordlist - $contextlength); $i++) {
    print OUTPUT $wordlist[$i] ;
    $contextlist[$i] = $wordlist[$i] ;
    for ($j=1; $j<$contextlength; $j++) {
        print OUTPUT " " . $wordlist[$i + $j] ;
        $contextlist[$i] = $contextlist[$i] . " " . $wordlist[$i + $j] ;
    }
    print OUTPUT "\n" ;
}
```

CONTEXTS
KWIC procedure: 3. Keyword List

```perl
@wordlist = split(/ /,$wordlist) ;
@sortedwordlist = sort { $a cmp $b } @wordlist ;
$prev = "" ;
$count = 0;
@uniquewordlist = () ;
for ( $i=0; $i<@sortedwordlist; $i++ ) {
    $a = $sortedwordlist[$i] ;
    if ( $a ne $prev ) {
        $prev = $a ;
        print OUTPUT $a . "\n" ;
        $uniquewordlist[$count] = $a ;
        $count++ ;
    }
}
```

KEYWORD LIST
KWIC procedure: 4. Search

for ($i=0; $i<@uniquewordlist; $i++) {
    $a = $uniquewordlist[$i] ;
    for ($j=0; $j<@contextlist; $j++) {
        @context = split(/ /,$contextlist[$j]) ;
        if ($a eq $context[2]) {
            print OUTPUT $context ;
        }
    }
}
KWIC procedure: 5. Format

1. Design a page layout with text objects:
   1. title
   2. headings
   3. body text
   4. tables

2. Implement - to test the algorithm - in HTML
KWIC procedure: Source

• The Perl implementation follows the procedure exactly
• However, the code is for demonstration purposes only, because it does not allow
  – flexible handling of contexts and filenames
  – treatment of more than one text
  – modularity of organisation
  – format scalability and search efficiency
• Project: re-write the code to do these things

PERL SOURCE CODE
KWIC: Scaling Up

• The Ibibio Concordance was made using exactly the same procedure, but:
  – using UNIX (Linux) shell scripting, not Perl
    • because this is the traditional method
    • today one could also use Python
  – using the LaTeX formatting language, not HTML
    • because this is much more flexible
    • the Toolbox system uses the RDF format for output:
      – Multi-Dictionary Formatter (MDF), or
      – Lexique Pro
    • today one could also use XML stylesheets

IBIBIO CONCORDANCE
QUIZ

• In which programming languages could the concordance software be implemented?
• What are the problems with the demonstration software which need to be removed in a later realistic project?
DICTIONARY MAKING
KWIC: Dictionary Making

• The function of a KWIC is
  – to make searching for lexical information more efficient by putting context information about words in one place
  – for making “Word Sketches” (Adam Kilgarriff)
    • grammatical descriptions: parts of speech
    • dictionaries: examples of use, collocations, ...

• Project: Make concordances from your text corpora and use them to collect lexical information for your Toolbox lexical databases
The Status of Dictionaries

• Remember that the dictionary is
  – one of the three main components of language documentation:
    • corpus of recordings and texts
    • dictionary
    • sketch grammar
  – the central component of any linguistic description
  – the most useful linguistic product for use by the speech community, or non-linguists in general
The Ibibio Dictionary

• The Ibibio Dictionary
  – uses information from Elaine Kaufmann's Ibibio Dictionary
  – the information was re-typed into an Office table format
  – this was converted into
    • Toolbox format for further lexicographic extension
    • LaTeX for formatting (cf. the Ibibio Concordance)

• Project: extend the Ibibio corpus, concordance, dictionary in scope & content
Conclusion

• It is faster to do this this way if you
  – have a large text corpus
  – want to make
    • a detailed syntactic or morphological description
    • a large dictionary
  – have less than a lifetime in which to do this
Conclusion

• It is faster to do this this way if you
  – have a large text corpus
  – want to make
    • a detailed syntactic or morphological description
    • a large dictionary
  – have less than a lifetime in which to do this
Homework

• Define a KWIC concordance
• Remember the KWIC construction diagramme
• Describe what is done at each stage
• Illustrate what is done at each stage, with examples from a short text
Conclusion

• It is much better to do this this way if you
  – have a large text corpus
  – want to make
    • a detailed syntactic or morphological description
    • a large dictionary
  – have less than a lifetime in which to do this

☺
Homework

• Remember the KWIC concordance construction diagramme.
• Describe the 6 stages of KWIC concordance construction.
• What can a KWIC concordance be used for?