22 June 2005

Lexicography: theory and practice

Dafydd Gibbon
gibbon@uni-bielefeld.de
Overview

- Why does everyone know what a dictionary is?
- Questions for the lexicographer (and for you)
- About dictionaries:
  - Types of dictionary
  - Lexemes, dictionary entries
  - How to define “definition”...
- Gestures as words?
- Some of the theory behind dictionaries:
  - megastructure
  - macrostructure
  - microstructure
  - mesostructure
- Practicalities
  - Lexica as tables
  - Corpus analysis
Why does everyone know what a dictionary is?

- “The dictionary” is what **everyone** relies on for
  - meanings, use, etymology, ...
  - Scrabble, spelling, pronunciation, ...
- Why?
  - A dictionary is the basic *inventory of fixed signs* of a language, lexical signs, together with all their properties.
- Lexical sign, lexeme:
  - **poodle** [puːdl] *n* A pet dog with a haircut. *Blair is Bush's poodle.*
- What is a sign?
  - A *form* (pronunciation & orthography) with a *meaning* and a *use*
  - The relation between *form* and *meaning* can be
    - symbolic: the usual case - “l'arbitraire du signe”
    - iconic: form-meaning similarity, e.g. onomatopoeia
    - indexical: dependence of meaning on the situational context of the form e.g. *I, you, this, that, here, now*, - “ego, hic et nunc”
Questions for the lexicographer (and for you)

• Overview:
  - Who or what is a lexicographer?
  - How does a lexicographer organise his work?
  - How do lexicography projects with publishers work?
  - Where does a lexicographer get his lexical information from?
  - Which tools does a lexicographer use to make a dictionary?
  - What kinds of dictionary are there?
  - Which parts does a dictionary typically have?

• Discussion points:
  - What kinds of dictionary are you familiar with?
  - What kinds of dictionary do you have at home?
  - What specific purposes do you use them for?
  - How good do they think they are?
  - Does “the dictionary” exist?
Dictionary entries - defining 'definition'

Longmans Dictionary of Contemporary English, 1987

**definitely** /ˌdefɪnɪˈtli/ adv 1 without doubt; clearly: *That was definitely the best play I’ve seen all year.* | *She is definitely coming/definitely not coming.* | *“Is she coming?” “Definitely!”* | *“Do you smoke?” “Definitely not!”* 2 in a definite way: *He explained his intentions very definitely.* —see certainly (USAGE)

**definition** /ˌdefɪnɪˈʃən/ n 1 [C] an exact statement of the meaning, nature, or limits of something, esp. of a word or phrase: *the definitions in a dictionary* | *An English person is British by definition.* (—it is part of the meaning of being “English”) 2 [U] clearness of shape, colour, or sound: *This photograph lacks definition.*

**definitive** /drɪˈfɪnɪtɪv/ adj 1 providing a last decision that cannot be questioned or changed; CONCLUSIVE: *a definitive decision by the supreme court* 2 that cannot be improved as a treatment of a particular subject: *She’s written the definitive life of Lord Byron.* —compare AU-
Definition chains

Definition by enumeration of hyponyms

Definition by exclusion of co-hyponym

genera proxima
Dictionary types

eczema /ˈɛksəmə/ 'eksəmə/ n [U] a red swollen condition of the skin

Edam /'iːdəm, -dæm/ n [U] a yellow cheese from the Netherlands made in balls covered with red wax

eddy /'edi/ n a circular movement of water, wind, dust, smoke, etc.: The little paper boat was caught in an eddy and spun round and round in the water.

eddy /ˈedi/ n [I] (of water, wind, dust, smoke, etc.) to move round and round: The mist eddied round the old house.

Eden /'iːdən/ n [ɪdə] the garden where according to the Bible Adam and Eve, the first human beings, lived before their disobedience to God, considered as a place of complete happiness

edge /ˈedʒ/ n 1 the part or place where something ends or begins or that is farthest from its centre: Don’t go too near the edge of the cliff. She stood by the water’s edge. If you don’t want it, leave it on the edge of your plate. Can you stand a coin up on its edge? (fig.) He felt he was on the edge of madness. 2 the thin sharp cutting part of
Dictionary types

Semasiological dictionary
(reader's dictionary, decoding dictionary)

Eczema /ɛkˈzɛmə/ n [U] a swollen condition of the skin.

Edam /ˈiːdəm, -dæm/ n [U] a yellow cheese from the Netherlands made in balls covered with red wax.

Eddy /ˈedi/ n a circular movement of water, wind, dust, smoke, etc.: *The little paper boat was caught in an eddy and spun round and round in the water.*

Eddy /ˈedi/ v [I] (of water, wind, dust, smoke, etc.) to move round and round: *The mist eddied round the old house.*

Eden /ˈiːdn/ n [the] the garden where according to the Bible Adam and Eve, the first human beings, lived before their disobedience to God, considered as a place of complete happiness.

Edge /ˈedʒ/ n 1 the part or place where something ends or begins or that is farthest from its centre: *Don’t go too near the edge of the cliff.* *She stood by the water’s edge.* *If you don’t want it, leave it on the edge of your plate.* *Can you stand a coin up on its edge?* (fig.) *He felt he was on the edge of madness.* 2 the thin sharp cutting part of
Dictionary types

Semasiological dictionary
(reader's dictionary, decoding dictionary)

Introduction to Applied Linguistics

Lexicography: Theory and Practice (Gibbon)
Dictionary types

Semasiological dictionary (reader's dictionary, decoding dictionary)

Onomasiological dictionary (writer's dictionary, encoding dictionary)
**Gestures as words?**

**Scenario:**

- fairy story
- semi-professional story-teller
- about a little donkey ...

---

<table>
<thead>
<tr>
<th>CoGes</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>die</td>
<td>wachsen</td>
</tr>
<tr>
<td>phrases</td>
<td>ihm</td>
<td>fein</td>
</tr>
<tr>
<td>function</td>
<td>hoch</td>
<td></td>
</tr>
<tr>
<td>lemma</td>
<td>und</td>
<td>gerade</td>
</tr>
<tr>
<td>frequency</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>tones</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>H^+</td>
<td></td>
</tr>
<tr>
<td>syllables</td>
<td>i</td>
<td>a</td>
</tr>
</tbody>
</table>

---

**Introduction to Applied Linguistics**

**Lexicography: Theory and Practice (Gibbon)**
Integrated lexicon theory

• Integrated Lexicon (ILEX):
  - to accommodate multimodal information
  - in a clear architecture

• Architecture for lexicographic development:
  - Layers of lexical abstraction
  - Lexicon structure:
    - megastructure
      • macrostructure (lexicon body) and metadata+blurb
    - macrostructure
      • organisation of lexical entries
    - mesostructure
      • generalisations based on DATCATS
    - microstructure
      • DATCATS, types of lexical information
A lexicon is a set of generalisations over

- **linguistic objects**
  - typical linguistic objects are phonemes, tones, morphs, words, phrases, gestures, dialogue contributions, texts
  - A linguistic object is a sign, i.e. an abstract object with two semantic mappings to representations of the world:
    - modality semantic mapping (phonetic, orthographic, ...)
    - object semantic mapping (as generally understood)
  - Linguistic objects are typically arranged in a rank hierarchy and enter into relations of temporal and spatial precedence and overlap.

- **identified**
  - the criterion identified means, in general, transcribed and annotated (time-stamped with respect to an acoustic or video speech recording, or space-stamped with respect to a document).

- **in a speech or text corpus**
  - a corpus is a set of speech recordings or documents (primary data)
  - with associated transcriptions, annotations (secondary data)
  - and metadata (tertiary data)
ILEX: layers of lexicon types

LEXICON

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

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

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

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

CORPUS

Secondary data
(transcription, annotation, metadata)

Primary data
(audio / video recording)

Introduction to Applied Linguistics

Lexicography: Theory and Practice (Gibbon)
ILEX: lexical megastructure

- Concept due to R.R.K. Hartmann

- Metadata
  - Title
  - Author
  - ...
  - (cf. Dublin Core specifications with OLAC and IMDI extensions)

- Lexicon body
  - macrostructure
  - mesostructure
  - microstructure
ILEX: lexical macrostructure

- Overall arrangement of lexical objects in Layers 1,..., n:
  - declarative
  - procedural: optimised for access
    - onomasiological, semasiological, ...
    - table (lemmata x properties)
    - tree (thesaurus)
    - other indexings/sortings by different microstructure elements
      - aka DATCATs
      - aka types of lexical information

- Ontology of Level 3 lexicon entry objects:
  - Headword (western lexicon tradition)
  - Stem (used in Arabic and Hebrew traditions)
  - Concept (termbanks, thesaurus)
  - Number of strokes (Japanese)
  - Inflected form (full form lexicon)
  - Problem with lexicalised noun class prefixes (Niger-Congo)
ILEX: lexical microstructure

- Lexical entry: traditionally, headword + article
- Structure of individual lexicon entries - DATCATS:
  - Orthography
  - Pronunciation
  - POS
  - Definition, etc.

- On procedural grounds
  - Field internal structure is often a shallow hierarchy (with some properties of mesostructure), e.g. sub-senses
  - For procedural purposes keys/names, sort orders, are required.
  - Otherwise headwords are redundant.

<table>
<thead>
<tr>
<th>ID</th>
<th>Orthography</th>
<th>Pronunciation</th>
<th>POS</th>
<th>Gender</th>
<th>Morphology</th>
<th>Inflection</th>
<th>Definition</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a donkey</td>
</tr>
</tbody>
</table>

Introduction to Applied Linguistics

Lexicography: Theory and Practice (Gibbon)
ILEX: central microstructure categories

Two points of connection to the world:

- **Content semantics**
  - classic semantic interpretation

- **Expression semantics**
  - interpretation function for
    - phonetic interpretation
    - text rendering:
      - layout
        - line
        - page
        - hypertext
      - font choice
      - highlights
  - relevance of
    - time models
    - space models
    - spatiotemporal models
ILEX: lexical mesostructure

Concept introduced in ILEX approach

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...

2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...

3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)

4. Corpus references (concordance)
ILEX: lexical mesostructure types

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...
2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...
3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...) 
4. Corpus references (concordance)
ILEX: lexical mesostructure types

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...
2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...
3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
4. Corpus references (concordance)
ILEX: lexical mesostructure types

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...
2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...
3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...) 
4. Corpus references (concordance)
ILEX: lexical mesostructure types

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...
2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...
3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
4. Corpus references (concordance)
ILEX: lexical mesostructure types

1. Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantic, ...
2. Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling, ...
3. Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
4. Corpus references (concordance)
ILEX lexicon architecture

1. MEGASTRUCTURE

2. METAINFORMATION
   Metadata, front matter, back matter

3. MACROSTRUCTURE

4. MICROSTRUCTURE
   (data categories, types of lexical information)

5. MESOSTRUCTURE
   Sketch grammar, i.e. conventions for generalisations over microstructure:
   - orthography
   - pronunciation
   - word formation
   - syntax
   - definitions
   - examples

Introduction to Applied Linguistics
Lexicography: Theory and Practice (Gibbon)
Practicalities: microstructure tables

Printed dictionary format

eczema /ˈɛksəmə/ n [U] a red swollen condition of the skin
Edam /ˈiːdəm, -dæm/ n [U] a yellow cheese from the Netherlands made in balls covered with red wax
eddy¹ /ˈedi/ n a circular movement of water, wind, dust, smoke, etc.: The little paper boat was caught in an eddy and spun round and round in the water.
eddy² v [I] (of water, wind, dust, smoke, etc.) to move round and round: The mist eddied round the old house.
Eden /ˈiːdən/ n [the] the garden where according to the Bible Adam and Eve, the first human beings, lived before their disobedience to God, considered as a place of complete happiness
edge¹ /ɛdʒ/ n 1 the part or place where something ends or begins or that is farthest from its centre: Don’t go too near the edge of the cliff. | She stood by the water’s edge. | If you don’t want it, leave it on the edge of your plate. | Can you stand a coin up on its edge? (fig.) He felt he was on the edge of madness. 2 the thin sharp cutting part of
Practicalities: microstructure tables

Printed dictionary format

Database format:

<table>
<thead>
<tr>
<th>Headword</th>
<th>Pronunciation</th>
<th>POS</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>eddy</td>
<td>eddi</td>
<td>n</td>
<td>a circular movement of water, wind, dust, smoke, etc.</td>
<td>The little paper boat was caught in an eddy and spun round and round in the water</td>
</tr>
</tbody>
</table>

Sentence examples:
- eddy: The little paper boat was caught in an eddy and spun round and round in the water.
- Eden: The garden where according to the Bible Adam and Eve, the first human beings, lived before their disobedience to God, considered as a place of complete happiness.
- edge: The thin sharp cutting part of.
Concordances

A concordance is a special kind of dictionary:
- each word in a text corpus is linked with its contexts of occurrence in this corpus
- (note: Google is a special form of concordance)

• Example:
- “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

• Simple KWIC concordance:
  1973 when i arrived
  a foggy march night
  arrived on the midnight
  calais was on a
  england from calais
  ferry sight of england
  first march night in calais
  foggy arrived on the
  from calais
  i 1973 when i
  in night in 1973
  march ferry from calais
  midnight first sight of
  in 1973 when
  england was on
  of a foggy march
  on the midnight ferry
  on england was
  sight midnight ferry from
  the on a foggy
  was when
  i arrived on
Quiz

• What is a *semasiological dictionary*?
• What information is typically found in a lexical entry?
• Describe a possible dictionary microstructure for *bread*.
• Make a KWIC concordance for the following text:
  “Once upon a time it snowed all year round.”
• What is the *macrostructure* of a dictionary?
• What is a *definition chain*?