The lexicon in computational linguistics

Ann Copestake
aac@csli.stanford.edu

Course overview

1. Introduction: applications and representation issues
2. Lexical acquisition: practical and theoretical problems
3. Lexical disambiguation and sense distinction
4. Towards more adequate models of the lexicon

1 Part I: applications and representation issues

1.1 Grammar checkers

Complex lexical knowledge is required:
Agreement phenomena:

(1) a * The dog are very happy
 b Our team are very happy that their contracts have been renewed
 c Our team of computational linguistics researchers are very happy that their contracts have been renewed
 d A forest of oak trees was damaged

Some nouns denoting groups of animate individuals can take plural agreement in some dialects (usual in British English). Similar things happen in AmE when 'of' phrase is present.

(2) a The total number of participants was fifty
 b A number of participants were very unhappy
 c A number of trees were damaged

Pseudo-partitives are slightly different.

Lexical requirement for this phenomenon Minimally, nouns which can take plural agreement when morphologically singular need to be distinguished in the lexicon of a grammar checker. Dialect differences must be allowed for.

1.2 Part-of-speech tagging

Given a corpus, tag the words in the text with the part-of-speech that is appropriate in that context.
For instance:

```
we can table that
we_PRO can_AUX table_V that_DEM
we_PRO can_N table_N that_DEM
```

Driven by transition probabilities: e.g., PRO followed by N is less likely than PRO followed by AUX. Normal approach: program takes a training corpus of words that have been manually tagged (maybe a few thousand) and learns most probable tags in a given context. (Systems can actually work without training data if there are enough unambiguous words.)

Lexical requirement A string is associated with all its possible tags.

```
table N, V
```

After tagging a corpus, lexical probabilities can be extracted. For instance:

```
table N 0.95, V 0.05
```

Lexical probabilities are the most important source of information in tagging. Approx. 90% accuracy just from choosing lexically most probable tag - approx 95% accuracy overall.

Unknown words In principle, all unknown words can be given all possible tags. Practically, turns out to be better to use morphological cues as much as possible.

1.3 Speech synthesis

Obvious requirement: pronunciation lexicon.
Other requirements for high quality synthesis include:

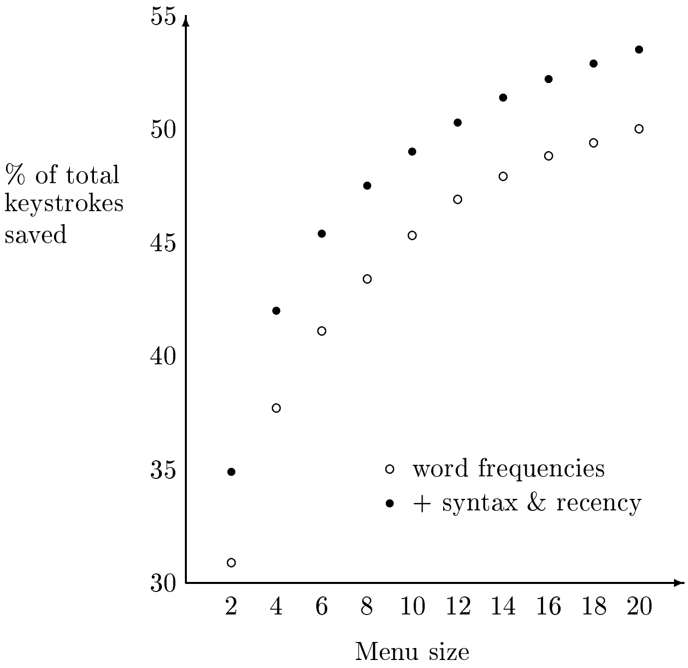
1. Disambiguation of homographs: e.g. *bow*, *1850*, *1491*
2. Phrasal stress: e.g. compound nouns
3. Intonation

1.4 Word and phrase prediction for text-to-speech for speech prosthesis

People who have lost speech due to physical disability may use a text-to-speech system to communicate, but need all the help they can get in speeding up text input.

- Present a menu of the most likely words and phrases (in context)
- User enters words letter-by-letter, choosing item from menu if available: e.g., after user chooses 't' — menu has words starting with 't' in probability order, after 't', 'a' — menu shows *table*, user chooses this and saves three keystrokes
- Most important thing is to learn lexicon from individual user
  - Menu size 8, around 44% keystroke saving simply on "words" and frequencies acquired from user.
  - Words and frequencies from text corpus: about 35% keystroke saving.
  - User frequencies plus part-of-speech transitions: about 47%.
- More elaborate information probably not very helpful for prediction — costs of elaboration (mainly lexical), diminishing returns.

Prediction results by menu size



1.5 Linguistically motivated parsing and generation systems

Many such systems: examples here are taken from the CSLI LinGO project, using HPSG.

Semantic representation is the goal

```
Browne interviewed Abrams on Thursday

(S
 (S (NP (NP (Browne)))
  (VP (VP (V (interviewed (V))) (NP (NP (Abrams))))
   (PP-I (P-I (P-I (on))) (N (N (Thursday)))))))

[ TOP: h1
  INDEX: e2: PAST*
  LISZT: <
          [ named_rel
            HANDEL: h3
            INST: x5
            NAMED: BROWNE]
          [ def_np_rel
            HANDEL: h6
            BV: x5
            RESTR: h8
```

```
            SCOPE: h9]
          [ _interview_v_rel
            HANDEL: h11
            EVENT: e2
            ARG1: x5
            ARG3: x13]
          [ named_rel
            HANDEL: h14
            INST: x13
            NAMED: ABRAMS]
          [ def_np_rel
            HANDEL: h16
            BV: x13
            RESTR: h18
            SCOPE: h19]
          [ _on_temp_rel
            HANDEL: h11
            ARG: e2
            ARG3: x22]
          [ dofw_rel
            HANDEL: h24
            INST: x22
            DAY: THU]
          [ def_np_rel
            HANDEL: h26
            BV: x22
            RESTR: h28
            SCOPE: h29]
          [ prpstn_rel
            HANDEL: h1
            SOA: h32]
          >
  H-CONS: <
            h8 =_q h3
            h18 =_q h14
            h28 =_q h24
            h32 =_q h11
          > ]
```

One of the scoped representations

```
prpstn(def_np(x5, named(x5, BROWNE), def_np(x13, named(x13, ABRAMS),
def_np(x22, dofw(x22, THU), _interview_v(e2, x5, x13) /\
_on_temp(e2, x22)))))
```

Lexical requirements

1. orthography
2. morphology
3. syntax
4. semantics
5. links between these
6. application-specific information (maybe)

Lexical entry

```
interview_v1 := mv_np*_trans_le &
  [ STEM < "interview" >,
    SYNSEM.LOCAL.KEYS.KEY _interview_v_rel ].

interview_n1 := intr_noun_word_le &
  [ STEM < "interview" >,
    SYNSEM.LOCAL.KEYS.KEY _interview_rel ].
```

Lexical sign (showing about 5% of the structure)

$$\left[\begin{array}{l}
\textbf{mv_np*_trans_le} \\
\text{STEM} : \left[\begin{array}{l} \textbf{*cons*} \\ \text{FIRST} : \textbf{interview} \\ \text{REST} : \textbf{*null*} \end{array}\right] \\
\text{SYNSEM} : \left[\begin{array}{l}
\textbf{np_trans_verb} \\
\text{LOCAL} : \left[\begin{array}{l}
\textbf{local} \\
\text{CAT} : \left[\begin{array}{l} \textbf{cat} \\ \text{HEAD} = \boxed{\textbf{verb*}} \\ \text{VAL} = \boxed{\textbf{valence}} \end{array}\right] \\
\text{CONT} = \boxed{\textbf{mrs}} \\
\text{KEYS} = \boxed{\textbf{keys}}
\end{array}\right] \\
\text{NON-LOCAL} = \boxed{\textbf{non-local}} \\
\text{LEX} : +
\end{array}\right] \\
\text{ARGS} = \boxed{\textbf{*cons*}} \\
\text{INFLECTED} : \textbf{-}
\end{array}\right]$$

Inheritance

```
mv_np*_trans_le := main_verb & reg_lex_entry &
  [ SYNSEM np_trans_verb &
   [ LOCAL.CAT.VAL.COMPS < [ OPT + ] > ] ].
```

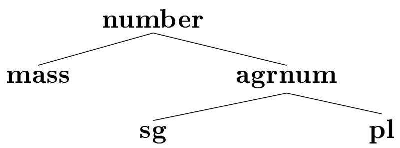
i.e. the type **mv_np*_trans_le** (where mv stands for main-verb) inherits from the types **main_verb** and **reg_lex_entry** and specifies its SYNSEM value as **np_trans_verb** and its complement is optional.

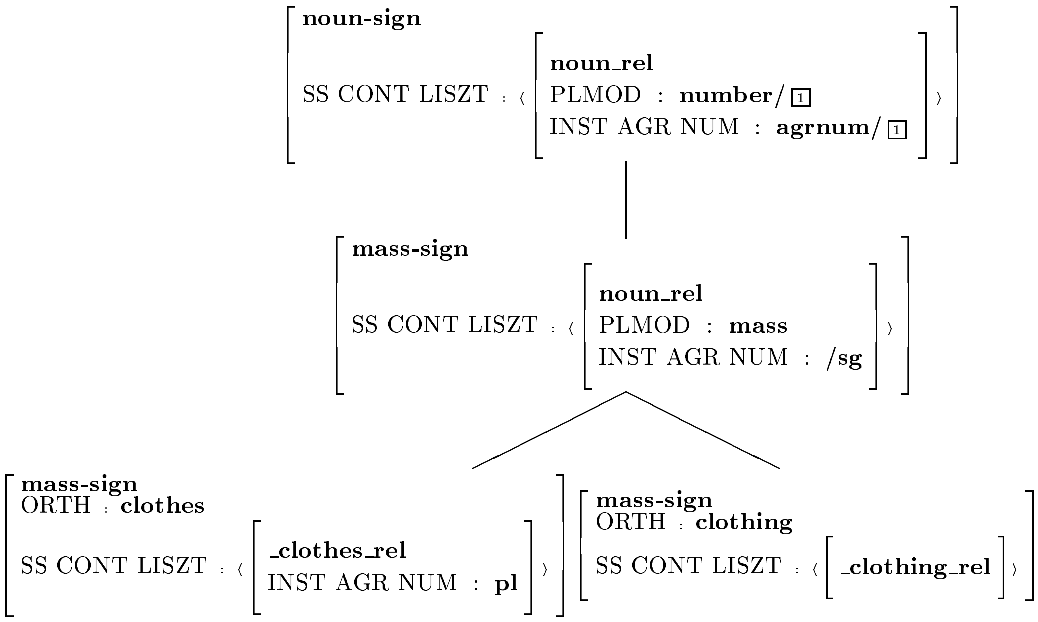
```
np_trans_verb := verb_synsem & np_trans_subst & trans_lt.
```

Lexical rules

1. ‘syntactic’ — e.g., subject auxiliary inversion
2. inflectional morphology
3. derivational morphology
4. verb alternations — e.g., dative
5. sense extensions — e.g., grinding

Default inheritance





$$\left[\begin{array}{l} \textbf{mass-sign} \\ \text{ORTH} : \textbf{clothes} \\ \text{SYNSEM CONT LISZT} : \left\langle \left[\begin{array}{l} \textbf{\_clothes\_rel} \\ \text{PLMOD} : \textbf{mass} \\ \text{INST AGR NUM} : \textbf{pl} \end{array}\right] \right\rangle \end{array}\right]$$

$$\left[\begin{array}{l} \textbf{mass-sign} \\ \text{ORTH} : \textbf{clothing} \\ \text{SYNSEM CONT LISZT} : \left\langle \left[\begin{array}{l} \textbf{\_clothing\_rel} \\ \text{PLMOD} : \textbf{mass} \\ \text{INST AGR NUM} : \textbf{sg} \end{array}\right] \right\rangle \end{array}\right]$$

Selectional restrictions

(3) a The book is on the table
  b The book is on the atom
  c Put the book on the atom
  d Put the book on the table
  e Devein the shrimp

If these were absolute, could easily encode with type hierarchy and unification or analogous mechanism.
But:

(4) a The electrons form a cloud round the nucleus of the atom
  b That's not far from my idea
  c Devein the worms

So better to encode using a preference / statistical mechanism. Old real-world vs lexical knowledge debate.

More 'linguistic' restrictions
We will meet **on** Thursday
We will meet **in** the morning
We will meet **on** Thursday morning
Choice of preposition should be dealt with in the grammar, because

  * We will meet in Thursday

(bad sentence in any context where Thursday refers to a day of the week). Contrast:

  We will meet in 48 hours

Distinction may be semantic, but not (crudely) denotational. Language specificity.
Temporal location prepositions: English generator given
e.g. temp-loc(e, x), day-of-week$(x, \text{Thursday})$, generates *on Thursday* because of restrictions in monolingual lexical entries.

1.6 The lexicon in machine translation

General model: two (or more) monolingual lexicons and a transfer lexicon.

Conditional statements in the transfer lexicon
"Easy" examples:

- *eat* $\leftrightarrow$ *essen* / *fressen* depends on whether subject is human or other animal (except can be used of human to imply animal-like eating)
- *worker* $\leftrightarrow$ *Arbeiter* / *Arbeiterin* depends on gender (might be extremely difficult to resolve)

Difficult examples:

- *kommen* $\leftrightarrow$ *come* / *get to* etc depends on complex factors about location and 'ownership' of space

  (5) You could come to the office.
  (5) * How do I come to the office?

- *Termin* $\leftrightarrow$ *meeting* / *appointment* depends on factors like relative social status of participants

Issues What should be encoded in the transfer lexicon? What should we use statistics for? What should the monolingual grammar do?

Lexical generalisations in Machine Translation
Verbs of directed motion, with terminated path (Talmy, e.g., 1986):

Kim swam across the river
$\text{swim}(e1, K), \text{across}(e1, x), \text{river}(x)$

Kim cruzó el río nadando
$\text{cruzar}(e2, K, x), \text{rio}(x), \text{WHILE}(e1, e2), \text{nadar}(e1, K)$

(here WHILE is intended to capture semantics of gerundive, after Sanfilippo, 1993).
Transfer rule required is:

$\text{swim}(e, w), \text{across}(e, z) \leftrightarrow$
$\text{nadar}(e, w), \text{cruzar}(e', w, z), \text{WHILE}(e, e')$

Generalise: assume PATH, MOVE-PATH, MOVE are (non-language specific) generalisations of the classes of predicate that include across, cruzar and swim, respectively.

$\text{swim}(e, w), \text{PATH}(e, z) \leftrightarrow \text{nadar}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$

$\text{MOVE}(e, w), \text{across}(e, z) \leftrightarrow \text{cruzar}(e', w, z)$

Metalevel patterns

$$\text{MOVE}(e, w) \leftrightarrow \text{MOVE}(e, w)$$
$$\Downarrow$$
$$\text{MOVE}(e, w), \text{PATH}(e, z) \leftrightarrow \text{MOVE}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$$

Thus, given $\text{swim}(e, w) \leftrightarrow \text{nadar}(e, w)$ we would also automatically get:
$\text{swim}(e, w), \text{PATH}(e, z) \leftrightarrow \text{nadar}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$

Compare monolingual lexical rules.
Could also use an interlingual notion of spatial prepositions and path components: e.g. semantics of *across* is $\text{CROSS-DIR}(e, z)$, semantics of *cruzar* is $\text{MOVE}(e, w) \wedge \text{CROSS-DIR}(e, z)$ and no separate transfer rule is necessary. Potentially plausible, because apparently limited range of spatial prepositions and types of path.

Language universals Engineering issue is how much grammar can be shared between languages, without unduly complicating the monolingual representation.

1.7 Summary

- many NLP applications — lexical requirements differ
- more elaborate systems at least need complex lexical organization, to capture generalizations (good linguistics and good engineering)
- has to be an underlying formalism for the representation, if only for reproducibility and maintainability
- always a problem of deciding how much to encode, especially when it comes to lexical semantics

2 PART II: Lexical acquisition

1. Hand coding
2. Reuse of existing resources
3. Machine readable dictionaries
4. Corpus-based approaches

2.1 Hand-coding

The traditional technique for full-blown, linguistically-motivated systems. For applications, hand-coding should be corpus-driven. Techniques exist to help the grammar engineer construct the lexicon, or to (partially) allow non-expert user to build lexicons.

Design principles for hand-coded lexicons

Minimize content in order to minimize acquisition problem.

Declarativity and human readability with compilation into a machine-friendly representation.

Modularity so components are reusable: e.g. distinct monolingual and transfer lexicons in an MT system.

Capture generalizations with inheritance (and lexical rules etc). Avoids errors, easier to maintain and expand.

Underspecification to reduce disambiguation for a particular application: e.g. *Zeitung* and *newspaper*.

2.2 Reusing resources

Primary problem is building **usable** lexicons: if it's usable then it's reusable!
But:

1. Procedural encoding (bad engineering?)
2. Lack of documentation, especially for semantics
3. Cost/benefit - not worth reusing a 100 word lexicon
4. Domain specificity
5. Legal issues

Many existing resources: wordlists (e.g. software companies), COMLEX, MRDs, Wordnet, EDR ...

2.3 Machine-readable dictionaries

1. Long history of research
2. Limited explicit syntactic information, except in English learners' dictionaries (OALD, LDOCE, Cobuild, CIDE)
3. Noun definitions can be analysed to derive taxonomies: inheritance hierarchies for semantic information.
4. Less published work on verbs (but e.g. Sanfilippo and Poznanski, 1992; Dorr et al,1995)
5. MindNet
6. Little published work on bilingual lexicons (Copestake et al, 1995, Klavans and Tzoukermann, 1996)

Deriving semantic information
Categorise lexical entries using predefined linguistically motivated classes.
E.g. automatically derived taxonomies.

Sauternes a type of sweet gold-coloured French wine

1. find *genus term* (*wine*)
2. disambiguate genus term wrt MRD senses
3. find the new entry's genus term ($drink^2$ 1) etc, etc
4. (possibly) refine classification using *differentia*

Taxonomy may be used directly, or:

1. associate taxonomy nodes with linguistic classes (e.g., encoded as types — e.g. **comestible_substance**)
2. information manually associated with type is inherited by lexical entries

MindNet

Bilingual dictionaries

MRDs: pros and cons
Advantages:

- dictionaries are used for manual lexicon construction
- broad coverage
- a labour-saving resource to augment a manually constructed core lexicon

Disadvantages:

- time-consuming to process initially

- highly variable quality
- different dictionaries require different strategies
- combining senses is non-trivial
- dictionaries inadequate even for **human** learners (ambiguities, little frequency information, obsolete or offensive terms, poor coverage of idioms)
- publishers' IPR
- restricted by printed medium

2.4 Corpus-based acquisition techniques

- Test corpus essential for hand-coded lexicons
- Semi-automatic acquisition techniques: range from simple use of concordances to post-filtering of automatically acquired material
- Automatic acquisition of monolingual information: taggers which cope with unknown words, subcategorisation frames (e.g. Brent, 1991, 1993; Manning, 1993, Ushioda *et al.*1993, Briscoe and Carroll, 1995) semantic (selectional) information (e.g. Ribas, 1995) etc
- Bilexicon acquisition from aligned corpora

Briscoe and Carroll

Corpora: pros and cons
Advantages:

- essential tool for human lexicographer
- domain-specific terminology and translations
- frequency information (some approaches)
- possibility of addressing ambiguity problem

Disadvantages:

- automatically processing corpora is a research challenge in itself
- large scale corpus may not exist for particular domain or language — parallel corpora especially difficult
- For MT: large scale results only demonstrated with one-to-one mappings, so far, automatic extraction unproven in full systems Also unclear how well results transfer to different classes of text

2.5 An example of acquisition: group nouns and pair nouns

Number agreement for English nouns: *group nouns* (e.g. *band*, *team*) are one class of exceptions to standard pattern, *pair nouns* (*scissors*, *trousers*, *binoculars*) are another.
Need to know about these exceptions for monolingual NLP and MT:

- Especially in BrE, many group nouns take sg or pl agreement
- In both AmE and BrE, pl anaphors are used to refer to sg group nouns
- Pair nouns require *pair of* as a classifier (**one scissors*, ?*three scissors*)
- Translation mismatches

Acquisition of group and pair nouns
Hand coding
Very inefficient unless done semi-automatically because vast majority of nouns are regular.
Existing resources
Thesaurus, WordNet useful for candidate group and pair nouns (on basis of semantic classes).
MRDs

- Pair nouns marked as obligatory plural in LDOCE (but marking is not unique to pair nouns).
- Group nouns marked in LDOCE syntax codes, but very inconsistently.
- Indirectly derive list of group nouns by looking at definition patterns (e.g. *group of* with human denoting noun).

  band: a group of musicians ...

Tagged corpora
Patterns such as sg noun followed by pl verb (filter for conjunctions, etc)

Rotherham_NP corporation_NN are_BER

Final remarks on acquisition

- Hand-coding, reuse, MRDs, corpora all have their advantages and disadvantages.
- Practical acquisition involves combination of techniques, making heavy use of machine-aided and semi-automatic methods.
- Test corpus is very important to measure progress, even if not used in any other way.

3 PART III: Disambiguation and sense distinction

3.1 Disambiguation

(6) They saw her duck

1. they visually perceived a female individual make a downward motion
2. they visually perceived an aquatic bird owned by a female individual
3. they cut with a sawing motion an aquatic bird owned by a female individual
4. they visually perceived a quantity of coarse cotton cloth owned by a female individual
5. ...

Harley and Glennon

Resnik and Yarowsky

Word uses and translation
Even in limited domains, words often have multiple translations: for example, (at least) five translations of *on* into French in ATIS (SRI, Cambridge):
fly to Boston on Delta
aller à Boston avec Delta

information on ground transport
des renseignements sur les transports publics

a meal on that flight
un repas à bord de ce vol

the aircraft which is used on this flight
l'avion qu'on utilise pour ce vol

leave on Monday
partir le lundi

3.2 What is a sense distinction anyway?

'Billiard ball' model of polysemy
Different senses correspond to distinct and unrelated lexical entries

bank 1 N $\lambda x[\text{bank1}(x)]$ (financial institution)

bank 2 N $\lambda x[\text{bank2}(x)]$ (mound of earth)

bank 3 N $\lambda x[\text{bank3}(x)]$ (building used by a bank1)

bank 4 ...

This model would predict we could identify discrete tight clusters of word usages. Reality is different ...

Dictionary definitions for *bank* — Collins English Dictionary (CED)
bank^2 *n.* **1** a long raised mass, esp. of earth; mound; ridge. **2** a slope, as of a hill. **3** the sloping side of any hollow in the ground, esp. when bordering a river. **4a** an elevated section rising to near the surface, of the bed of a sea, lake or river **b** *(in combination): sandbank, mudbank.*... *vb* **9** (when transitive, often followed by *up*) to form into a bank or mound ...
Note — definitions have a lot of structure (some of it implicit) — Vossen et al (1989)

Dictionary definitions for *bank* — Cambridge International Dictionary of English (CIDE)
bank RAISED GROUND *n* sloping raised land, esp. along the sides of a river, or a pile or mass of earth, clouds, etc

Dictionary senses are partly arbitrary — sense splitting techniques depend on editorial style (Atkins and Levin, 1992; Kilgarriff, 1992) — senses are (partly) ways of showing different conventional uses (CIDE does this with examples)

Problems with 'billiard ball' model

- Relatedness of senses is not captured (meaning postulates, but missing generalizations, unrestricted, etc)
- Productive extensions of sense (e.g., *bank* used to mean person/people in a bank)
- Arbitrariness of fine-grained sense distinctions
- Homonymy (i.e., total unrelatedness of sense) is fairly rare (rough estimate — 1 in 30 word forms): polysemy is ubiquitous

Linguistic notion of sense: ambiguity vs vagueness
ambiguity: senses can never be simultaneously used (without a pun): distinct lexical structures
vagueness: use is underspecified (with respect to some characteristic): single lexical structure
Standard formal accounts assume that only a single lexical structure can be involved in any interpretation.
(Vagueness is sometimes relevant linguistically, but often not: *glove* is vague between left- and right- handed glove, but no need to represent this explicitly.)

Ambiguity tests
is copredication possible?

(7) This book is 1000 pages long and totally unreadable

is there a pun?

(8) Fred and his driving licence both expired on Tuesday

can plural refer to a mixed collection?

(9) The terrorists blew up three banks

(10) The salesman sold me three brushes

Problems with ambiguity tests
Conflicting results — is *book* vague or ambiguous?

(11) John Updike has sold three books this year

False positives with puns:

(12) Rembrandt and our old janitor were both geniuses with a brush

Sentences can be ambiguous (intuitively) without lexical or syntactic ambiguity:

(13) Have you put any oil in the BMW?

- Is there a real distinction between vagueness and ambiguity?
  Claim here: useful to talk about lexical ambiguity when copredication is never possible.
- How to account for puns and copredication?
  Discourse can set up a context where there's a contrast, even when there's no ambiguity — e.g. contrast between Rembrandt and the janitor (Lascarides et al, 1996)

3.3 Polysemy: some distinctions

homonymy unrelated senses
bank:financial institution vs *bank*:mound

unpredictable polysemy unpredictable distinct senses
bank:tilt of plane vs *bank*:mound

sense extension general pattern of ambiguity
bank:to construct a *bank*:mound (also *heap*, *pile*, *mound*, *stack* etc)

constructional polysemy single structure in the lexicon, uses of this sense are sometimes differentiated by syntactic context
bank:mound (normally earth etc) vs e.g. *bank of cloud*
bank:building used by *bank*:financial institution

sense modulation single structure in the lexicon, different uses aren't morphologically or syntactically significant
bank:mound vs *bank*:river bank

(14) Removal of vegetation may destabilize banks

Polysemy: representation

homonymy and unpredictable polysemy distinct entries
(though note that collocates may be similar for unpredictable polysemy — e.g., *steep bank*)

sense extension lexical rules

constructional polysemy underspecification (maybe with defaults)

sense modulation probably not represented (i.e., sense is general, specialisations aren't represented at all)

Lexical rules

- Conventionally used for verb alternations, (derivational) morphology
- Map one sense into another
- Allow for morphological, syntactic and semantic effects

(15) Kim banked (up) the earth
$\text{bank}(x) \wedge \text{MADE-OF}(x, earth) \wedge \text{CONSTRUCT}(k, x)$

Note: such uses of lexical rules require

1. fairly detailed lexical semantic distinctions
2. results of productive application must have constructed semantics

Constructional polysemy

- A single sense assigned to a lexical entry is contextually specialised
- Implemented by underspecification
- Lexical entries may be multi-faceted
- Default information may be contextually or pragmatically overridden

Term *constructional polysemy* from Copestake and Briscoe (1995), but original concept partly based on ideas from Pustejovsky's Generative Lexicon framework.

(16) Mary enjoyed the book.
Mary enjoyed reading the book.

The apparent generalisation about *enjoy* and similar verbs is that semantically they always take eventualities. When the syntactic complement is an object-denoting NP, the sentence is nevertheless interpreted to mean that an event is enjoyed. Context can affect the interpretation: suppose we know or find out that Mary is a goat ...
drink has an intransitive use that implies a narrow-scope, existential drinkee, but default seems to be alcoholic drink (also *eat* and *bake*).

3.4 A statistical approach to polysemy: Schütze (1996)

- Senses as clusters
- 'Semantic space' is multidimensional, depending on association with other words (word senses)
- Gradient rather than discrete notion of sense: cline with homonymy as one extreme.
- Disambiguation — usages can be associated with sense clusters (model can handle intermediate cases).

Some issues

- Word association seems to be a good way of thinking about meaning (certainly given current state of the art in AI).
- Collocation (whether interpreted statistically or not), combined with frequency information, is the most effective disambiguation technique. (Even systems which do more complex reasoning for some disambiguation rely on collocation for majority of cases.)
- Some symbolic representation seems required for connection to syntax. Possibly also need some cut-off points on the cline — there do seem to be incompatible senses.
- Regularities — allows for acquisition (but Schütze mentions that it's possible that something more like lexical rules is the 'adult' representation).

Mixed statistical and symbolic approaches: sense frequency
Word senses often have widely different frequencies:
diet — first two WordNet senses

1. (a prescribed selection of foods)
   => fare – (the food and drink that are regularly consumed)
2. => legislature, legislative assembly, general assembly, law-makers

Intuition: *diet* 1 is much more frequent (not just because of denotation).
British National Corpus (BNC): *diet* 1 has probability of about 0.9, *diet* 2 about 0.005

Also statistical measure of productivity for generative processes. (lexical rules)

Collocation

roast	**bake**
beef, pork, chicken, head of lamb, camel, goat, !cow, lizard, crab, *ham, *gammon	bread, cake, pies
potato, chesnuts, turnip, crab apples, ?apple	ham, fish, snake, barracuda steak
coffee beans	apple, potato
duke	Gratin Dauphinois, pork chops with apples
metal ore	clay, concrete, earth
	pottery, resistance films

3.5 A philosophy

1. Sense distinction and semantic representation in the computational lexicon should be primarily driven by criteria involving syntax and morphology (phonology)

2. Making sense distinctions, and relating senses (by lexical rule or inheritance) is essential to capture generalisations

3. Some lexical semantic representation is essential to capture generalisations (possible more than most people think ...)

4. If sense distinctions don't affect processing in a particular component of a system, those distinctions shouldn't have to be made in that component — i.e., use underspecification, with specialisation as needed (compare scope underspecification)

5. Practically, for broad coverage systems, a more detailed meaning representation (for open-class words) must be approximated by statistical information about word relationships

4 PART IV: Towards more adequate models of the lexicon (maybe)

4.1 Generativity and semi-productivity

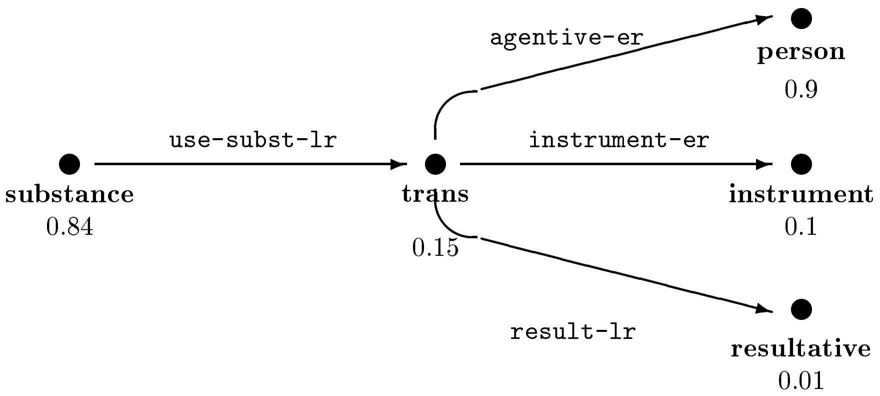
The computational linguist has to allow for productive word-formation (*fax*, *skateboard* as a verb etc, *faxable*) — we can never list all words of a language.
But semi-productivity and sense-proliferation . . . Forms which would arise if the alternations given by Levin (1992:267) are applied to *helicopter*:

(17) a The pilot helicoptered
b The pilot helicoptered over the forests
c Mrs Clinton was helicoptered to the base
d The pilot helicoptered the forests (cf they skated the canals)
e The pilot helicoptered his passengers sick

Levin doesn't claim these can all apply to all verbs in the class, and judgements of the grammaticality of such examples differ. The point is that even when such uses are plausible and attested, they are rare for the great majority of nouns which could in principle undergo the conversion.

$$\text{Prob(lexical-entry} \mid \text{word-form)} = \frac{freq\text{(lexical-entry with word-form)}}{freq\text{(word-form)}}$$

Probabilistic account



Representation of lacquer as a finite state machine with conditional probabilities that base word form (*lacquer*, *lacquerer*) is associated with a particular structure.

Productivity Derivational processes in general have very different productivities, considered as the proportion of possible inputs they actually apply to e.g. *er*-suffixation (*teacher*) vs. *eer*-suffixation (*profiteer*)

Paint nouns

word	BNC			CIDE		
	noun	verb	-er	noun	verb	-er
acrylic	104	0	0	+	-	-
calcimine	0	0	0	-	-	-
distemper	37	4	0	+	-	-
dye	291	153	19	+	+	-
emulsion	117	2	0	+	-	-
enamel	287	25	1	+	+	-
gouache	85	0	0	+	-	-
japan	17	1	0	-	-	-
lacquer	132	25	1	+	+	-
paint	1783	2170	965	+	+	+
primer	236	0	0	+	-	-
semigloss	0	0	0	-	-	-
shellac	35	1	0	-	-	-
tempera	29	0	0	+	-	-
tint	60	46	0	+	+	-
undercoat	44	0	0	+	-	-
varnish	231	37	1	+	+	-
veneer	156	12	0	+	-	-
watercolour	272	0	0	+	-	-

Class productivity estimates for noun-verb conversion

$$\text{Prod(lexical-rule)} = \frac{M}{N}$$

(where N is the number of attested lexical entries which match the lexical rule input and M is the number of attested output entries)

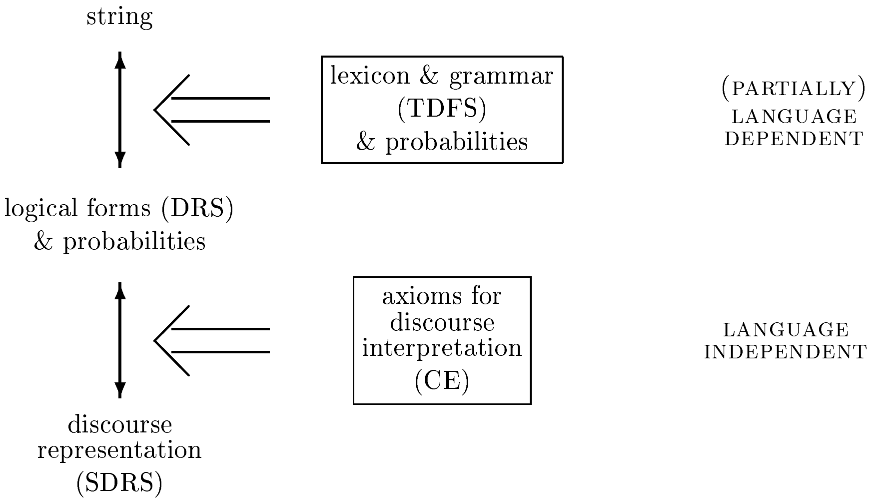
class	total	alternating	productivity
land vehicle	75	12	0.16
dance	41	9	0.22
paint	23	15	0.65
club	35	18	0.51

These can be used to estimate probabilities for unseen forms.

4.2 The lexicon and pragmatics

1. Lexical representation must integrate with syntax and compositional semantics
2. Lexicon should not allow arbitrary inference, lexical semantic information is not open-ended
3. Pragmatics should not access language-specific (conventional) information

Architecture



TDFS Typed Default Feature Structures

CE Commonsense Entailment

SRDS Segmented Discourse Representation Structures

Radical pragmatics and compounds

1. Compounds cannot all be conventionally/compositionally associated with a limited number of specific meanings (not generally possible to e.g., recover a verb)

2. Rules do not allow for novel types of usage

   (18) Please sit in the apple juice seat (Downing, 1977)

3. Reasoning about discourse situation is needed to resolve referent for novel cases

4. Since pragmatics is needed for interpretation of some compounds, it is unnecessary to assume any contribution of the lexicon in general

Cf Nunberg (1978) — *ham sandwich*,
Clark and Clark (1979) — *teapotting*

Complete underspecification of compound meaning
Rule for right-headed noun-noun compounds:

$$\begin{array}{cccc} \text{N0} & \longrightarrow & \text{N1} & \text{N2} \\ \lambda x[P(x) \wedge Q(y) \wedge R(x,y)] & & \lambda y[Q(y)] & \lambda x[P(x)] \end{array}$$

juice seat: $\lambda x[\text{seat}(x) \wedge \text{juice}(y) \wedge R(x,y)]$
R is underspecified, to be filled in by pragmatics, based on the discourse context.
(Here and below, we ignore the issue of the quantification of non-head element.)

Problems

1. Overgeneration — some classes of 'real-world-possible' compounds do not occur
2. Grammar of compounds ('possessive' compounds, stress assignment)
3. Conventional uses (not just a matter of listing lexicalized compounds — *used bookstore*, ?*jet blackbird*)
4. Failure to account for usual interpretations — most compounds actually do fit into small number of classes (e.g. Levi 1978)
5. Pragmatics does not have enough to work on — rule would predict that unseen compounds could only be used/interpreted in contexts which could disambiguate them

Overgeneration: German compounds with non-compound translations

Arzttermin	*doctor appointment
Terminvorschlag	* date proposal
Terminvereinbarung	* date agreement
Januarhälfte	* January half
Frühlingsanfang	* spring beginning

**doctor appointment/ doctor's appointment* — possessive compounds
**date agreement/ agreement on a date* — head derived from a PP-taking verb (*water seeker* vs **water looker*)
**spring beginning/ beginning of spring* — relational nouns as heads?

Possessive compounds
Many compounds with human-denoting first elements require a possessive in English.
blacksmith's hammer/ *blacksmith hammer* 'hammer of a type conventionally associated with a blacksmith' (also *driver's cab*, *widow's allowance* etc).
Not the usual possessive: compare (((his blacksmith)'s) hammer) with (his (blacksmith's hammer))
Adjective placement: *three English blacksmith's hammers/ *three blacksmith's English hammers.*
May be an agentivity requirement: *child seat*

Stress assignment
Compounds generally have leftmost stress (simplifying — see e.g., Bauer (1983), Liberman and Sproat (1992) for more accurate statement)
But compounds which have the interpretation 'Y made of X' (e.g., *nylon rope*, *oak table*) generally have main stress on the righthand noun.
Stress sometimes disambiguates:

(19) Sandy put the shirts into the cotton bag

with righthand stress — *bag made of cotton*
with leftmost stress — *bag for cotton* (other reading may still be available)

Conventional uses

1. Established compounds may be transparent but not predictable, are often dialect specific: *center divide*, *dustbin man* (vs. *rubbish man*)
2. Established compounds must be recorded somehow in the lexicon, at least for generation
3. Therefore, classes should also be lexical, to allow generalizations to be expressed
4. BUT: in context, even established compounds have alternative interpretations (Bauer's *garbage man* by analogy with *snowman*)

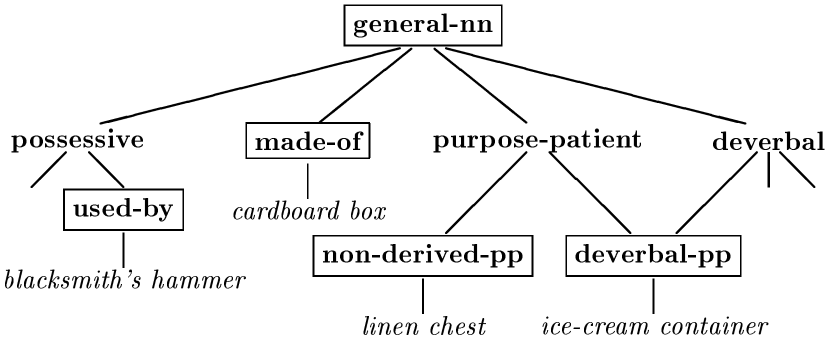
Principles of compound representation

- Compound meaning cannot be completely underspecified — the grammar/lexicon delimits the range of compounds, lists established compounds and lists conventional patterns.
- Patterns of interpretation have differing productivities.
- Some compounds are only possible in context (e.g., Downing (1977) *apple-juice seat*).
- Non-conventional contextual interpretations are available for all compounds in context (e.g. Bauer (1983) *garbage man* like *snowman*), but are much less likely than conventional readings.

Representing compounds

1. Hierarchy of noun-noun compound schemata represented in typed default feature structures (cf., grammar rules)
2. Schemata have associated productivity measurements
3. The most general schema is a catch-all — a completely underspecified interpretation to be resolved by pragmatics
4. General schema excludes more specific subschemata interpretations (so some *interpretations* of compounds are excluded)
5. Established compounds may have idiosyncratic interpretations or inherit from one or more schemata
6. Established compounds also have unestablished interpretations but these have minimal probabilities

Fragment of hierarchy of noun-noun compound schemata



boxed nodes are schemata
unboxed nodes just capture generalizations

Sample schemata

general-nn	N0	->	N1	N2
	$\lambda x[P(x) \wedge Q(y) \wedge R_c(x,y)]$		$\lambda y[Q(y)]$	$\lambda x[P(x)]$
	$\mathrm{R}_c = /?$		anything /stressed	anything
made-of	R_c = made-of		substance	physobj /stressed
purpose-patient	R_c = TELIC(N2)		anything	artifact

/ indicates that the value to its right is default information
All schemata define a value for R_c, the semantic relationship between parts of the compound
R_c predicates can often be reused from other parts of semantic description:
made-of — material constituency, e.g., as in Link (1983) on mass terms
telic role (Pustejovsky, 1995) encodes the purpose of an artifact (cf Johnston and Busa, 1995)

Interpretations of ***cotton bag***

1. $\lambda x[\text{cotton}(y) \wedge \text{bag}(x) \wedge \text{made-of}(y,x)]$
2. $\lambda x[\text{cotton}(y) \wedge \text{bag}(x) \wedge \text{TELIC}(\text{bag})(y,x)] = \lambda x[\text{cotton}(y) \wedge \text{bag}(x) \wedge \text{contain}(y,x)]$
3. $\lambda x[R_c(y,x) \wedge R_c =? \wedge \neg(\text{made-of}(y,x) \vee \text{contain}(y,x) \vee \ldots)]$

These interpretations are not all equally probable ...

Probabilities
In general, word senses often have widely different frequencies. e.g., *diet* — first two WordNet senses

1. (a prescribed selection of foods)
   => fare – (the food and drink that are regularly consumed)
2. => legislature, legislative assembly, general assembly, law-makers

diet 1 has probability of about 0.9, *diet* 2 about 0.005 (in the BNC)
Hypothesis: pragmatics doesn't consider infrequent senses unless most likely interpretation is incoherent
Estimation of probabilities for established uses is straightforward (if practically tedious), but unseen uses must be estimated via productivity estimate

Productivity of compound schemata

non-productive e.g., verb-object: *pickpocket*, *cutpurse*, but no recent examples (*bag-snatcher* not **snatchbag*)

moderate productivity e.g., has-part: *4-door car* but **sunroof car*

highly productive e.g., made-of: *ebony tripod* (though obviously not e.g., *juice chair*)

Productivity and probabilities

$$\text{Prod(cmp-schema)} = \frac{M+1}{N}$$

where N is the number of pairs of senses which match the schema input and M is the number of attested two-noun output forms in a corpus. Compounds with more than two nouns are ignored for simplicity. (cf., Briscoe and Copestake (1996) on semi-productivity in morphology and sense-extension)

Probabilities for unseen compounds

$\text{Unseen-prob-mass(cmp-form)} = \frac{\text{number-of-applicable-schemata(cmp-form)}}{\textit{freq}\text{(cmp-form)+number-of-applicable-schemata(cmp-form)}}$
Estimated-freq(interpretation$_i$ with cmp-form$_j$) =
Unseen-prob-mass(cmp-form$_j$)$\times \frac{Prod(cs_i)}{\sum Prod(cs_1),\ldots,Prod(cs_n)}$
(where $cs_1 \ldots cs_n$ are the compound schemata needed to derive the n unattested entries for the form$_j$)

cotton bag examples

(20) a Mary put a skirt in a cotton bag

$$\begin{array}{|l|}\hline e, x, y, z, w, t, now \\ \hline mary(x),\ skirt(y),\ cotton(w), \\ bag(z),\ put(e, x, y, z), \\ made\text{-}of(z, w) \\ \hline \end{array} \qquad P = 0.84$$

$$\begin{array}{|l|}\hline e, x, y, z, w, t, now \\ \hline mary(x),\ skirt(y),\ cotton(w), \\ bag(z),\ put(e, x, y, z), \\ contain(z, w) \\ \hline \end{array} \qquad P = 0.14$$

$$\begin{array}{|l|}\hline e, x, y, z, w, t, now \\ \hline mary(x),\ skirt(y),\ cotton(w), \\ bag(z),\ put(e, x, y, z), \\ R_c(z, w), R_c = ?, \\ \neg(made\text{-}of(z, w) \vee \\ contain(z, w) \vee \ldots) \\ \hline \end{array} \qquad P = 0.02$$

probabilities based on productivity figures for fabric/container compounds in the BNC, using WordNet for semantic categories

The role of pragmatics and discourse context in compound interpretation

1. accept most probable LF that is coherent in a discourse context
2. instantiate underspecified relationships (to the extent necessary)

(21) Mary sorted her laundry into various large bags.
She put her skirt in the cotton bag.

Lexically most likely interpretation is *bag made of cotton*: nothing makes this incoherent, so accepted.

(22) Mary sorted her laundry into various bags made from plastic.
She put her skirt into the cotton bag.

Lexically most likely interpretation is incoherent because the definite description cannot be accommodated into the discourse context.
Second-most likely interpretation — *bag used for cotton* — is OK

Formalizing discourse context

Segmented Discourse Representation Theory: SDRT (Asher, 1993) and Discourse in Commonsense Entailment:DICE (Lascarides and Asher, 1993).
In SDRT discourse is represented as a set of DRSs linked together with rhetorical relations such as *Elaboration* and *Contrast*
DICE specifies how background knowledge resources interact to determine rhetorical relations
Informally $Update(\tau, \alpha, \beta)$ is the SDRS with the discourse context τ, the new information β, which is attached via a rhetorical relation R to α, according to coherence conditions on R

SDRS S is well-defined (written $\downarrow S$) if there are no unresolved anaphora and if all constituents are attached with a rhetorical relation

Summary

- Knowledge of language includes knowledge of frequencies of word senses and of degrees of productivity
- Our approach allows weird compounds (*apple juice seat*) and weird interpretations (*garbage man* like *snowman*) in right discourse context, but very low probability
- Low probability meanings are only considered when higher probability meanings fail
- Similar approach to sense extension (*ham sandwich* etc)
- More work needed on distinction between novel and established compounds — context licensing examples such as *apple juice seat*, noteworthiness (Nunberg, 1995)

Relevance to practical NLP

- Discourse interpretation using CE is not practical for broad coverage systems (AI-complete)
- But use of probabilities reduces role of pragmatics — probabilities are partly conventional, partly real-world knowledge
- Full pragmatic component not necessary for systems which aid rather than replace a human e.g., report generation, augmentative and alternative communication (AAC), some approaches to MT
- Approach is consistent with philosophy of underspecification described in the last session. We can produce a general meaning that is specialized as needed.
- Questions: how should lexical probabilities be combined?
are lexical probabilities from large corpora generally useful?
what is the right granularity for senses?