

Natural Language Processing & Applications Notes on Selected Exercises

Phones and Phonemes

1. a) There are 3 phonemes in *Keith*. In IPA *Keith* is /kiθ/. Each can be independently substituted. For example:

/kiθ/ – /k/ + /h/ → /hiθ/ i.e. *heath*

/kiθ/ – /i/ + /ɪ/ → /kɪθ/ i.e. *kith*

/kiθ/ – /θ/ + /p/ → /kip/ i.e. *keep*

- b) There are 4 phonemes in *coughs*, which is pronounced [kʊfs] in SEE. Each can be independently substituted. For example:

/kʊfs/ – /k/ + /d/ → /dʊfs/ i.e. *doffs* (as in *he doffs his hat*).

/kʊfs/ – /ʊ/ + /ʌ/ → /kʌfs/ i.e. *cuffs*

/kʊfs/ – /f/ + /t/ → /kʊts/ i.e. *cots*

/kʊfs/ – /s/ + /i/ → /kʊfi/ i.e. *coffee*

3. I think that the ‘nasal assimilation’ rule does apply, at least in fast speech in SEE. (But note that it does not apply to all dialects of English, so your answer may well be different for your speech). English spelling shows the shift from /n/ to [m], but not that from /n/ to [ŋ] since there is no letter for this phone. Some examples:

in+polite /np/ → [mp] *impolite*

in+balance /nb/ → [mb] *imbalance*

in+tolerant /nt/ → [nt] *intolerant*

in+decisive /nd/ → [nd] *indecisive*

in+considerate /nk/ → [ŋk] *inconsiderate*

in+gratitude /ng/ → [ŋg] *ingratitude*

4. The rule for *Aston* and *Asda* can be written in several ways, depending on the level of generality required. Three possible rules are:

/s/ in the context ‘_ voiced alveolar stop’ becomes [z]

/s/ in the context ‘_ alveolar stop’ takes on the voicing of the stop

/s/ in the context ‘_ stop’ takes on the voicing of the stop

To test the last rule we need to find all combinations of /s/ + stop. Some examples:

clasp /sp/ → [sp]

husband /sb/ → [zb]

Aston /st/ → [st]

Asda /sd/ → [zd]

task /sk/ → [sk]

Asgard /sg/ → [zg] (Asgard was the home of the Norse gods.)

In ‘feature set’ notation the rule can be written in either of the forms below:

/s/ {stop,Voicing} → {fricative,alveolar,Voicing} {stop,Voicing}

/s/ → {fricative,alveolar,Voicing} : _ {stop,Voicing}

6. A possible rule for German is that alveolar stops are ‘de-voiced’ in the context ‘_ end-of-word’, i.e.

{stop,alveolar} → {stop,alveolar,voiceless} : _ end-of-word

In fact this rule works for all stops, not just alveolar stops (although other rules may operate to change the final output). Thus *Tag* (day) is pronounced [tak], whereas the plural *Tage* is pronounced [tagə]. How would you expect *gelb* (yellow) and *gelbe* to be pronounced? Many Germans when speaking English say, e.g., [hæf] for *have*. What does this suggest about the rule above?

7. Using only the nasal assimilation rule, we expect there to be 6 possible nasal+stop endings to English words. In spelling, *n* will appear where the sound is [ŋ]. The first column below shows the 6 expected endings; the second column some words whose spelling matches these endings. However, the actual pronunciation in SEE is as in the final column: [b] and [g] are deleted in the context ‘nasal _ end-of-word\OVAL(.,\D\F01()’ (Elsewhere they may or may not be deleted, e.g. *amber* is pronounced [æmbə], but in *lambing* the *b* is still silent.)

/mp/	<i>limp</i>	→	[mp]
/mb/	<i>limb</i>	→	[m]
/nt/	<i>lent</i>	→	[nt]
/nd/	<i>lend</i>	→	[nd]
/ŋk/	<i>rink</i>	→	[nk]
/ŋg/	<i>ring</i>	→	[ŋ]

In the feature set notation developed here:

/b/ → nothing : {nasal} _ end-of-word
 /g/ → nothing : {nasal} _ end-of-word

(Linguists tend to dislike rules containing specific phonemes, so we might ask what features /b/ and /g/ have that make this rule apply to them and not to /d/. I can’t see a more general rule – can you? Note also that the /ŋg/ → [ŋ] rule doesn’t operate in some dialects, including Birmingham English: see below.)

When this rule is combined with the nasal assimilation rule, one interesting output is obtained by treating the grapheme *ng* as two phonemes /ng/. Assimilation first changes /ng/ to [ŋg], after which the deletion rule removes the [g] to give just [ŋ].

This leads to an important difference between computing scientists and linguists as to the STATUS of phonological (and other) rules. Traditionally, many linguists have tried to find ‘the’ rules, assuming there is some underlying psycholinguistic reality. In the example here, there appear to be two possibilities for English:

1. Accept /ŋ/ as a phoneme. *Ring* will be represented as /rɪŋ/ and the ‘g deletion rule’ isn’t needed.
2. Reject /ŋ/ as a phoneme, and always use /ng/. Words containing terminal [ŋ] will be derived via /ng/ → /ŋg/ → [ŋ].

As computer scientists, we need only ask which rule(s) are more convenient, NOT which, if any, have some underlying reality. This means that we may be guided by the linguists’ theorizing, but are always free to adopt a more ad hoc approach. One advantage of (2) above is that there are some dialects of English in which the process seems to stop half-way. For example, *Birmingham* is pronounced in the local dialect roughly as [bə:mɪŋgəm]. Mapping spoken words in this dialect back to phonemes will be easier with (2). (1) would need a ‘g insertion’ rule in the forward direction.

8. The examples suggest that two rules are operating. One is that the /n/ is deleted unless before a vowel or a voiceless stop. This is not easy to write in the formal rule format adopted here:

/ðe/ /n/ not({vowel}|{stop,voiceless}) → /ðe/ not({vowel}|{stop,voiceless})

The other rule is that assimilation occurs between the /n/ (which is always voiced) and the voiceless stop, causing them both to be articulated at the same position in the mouth and voiced:

/ðe/ {nasal,alveolar,voiced} {stop,Posn,voiceless} →
 /ðe/ {nasal,Posn,voiced} {stop,Posn,voiced}

Notice that this rule means that unlike English, the phonological rules for Modern Greek will not form a context-sensitive phrase-structure grammar because two items change.

(It would be surprising if this rule applied only to the sequence /ðen/, but no evidence was presented in the question to generalize it. In fact, the rule does apply more generally. For example, to say ‘to the’ followed by a masculine noun we use /ston/ with the same rules.)

9. In feature set notation, a rule for *Tomkin* / *Tompkin* is:

$$\{\text{nasal, bilabial, voiced}\} \rightarrow \{\text{nasal, bilabial, voiced}\} \{\text{stop, bilabial, voiceless}\} \\ : _ \{\text{stop, non-bilabial, voiceless}\} :$$

This ensures that the sequence of phones changes from:

$$\{\text{nasal, bilabial, voiced}\} \{\text{stop, non-bilabial, voiceless}\}$$

where all three features are different between adjacent phones, to:

$$\{\text{nasal, bilabial, voiced}\} \{\text{stop, bilabial, voiceless}\} \{\text{stop, non-bilabial, voiceless}\}$$

where at most two features differ between phones.

To generalize to fricatives, we can simply duplicate the rule, substituting fricative for stop and labiodental for bilabial where appropriate, or write something like:

$$\{\text{nasal, labial, voiced}\} \rightarrow \{\text{nasal, labial, voiced}\} \{\text{stop|fricative, labial, voiceless}\} \\ : _ \{\text{stop|fricative, non-labial, voiceless}\} :$$

With this definition, *Tomson* yields *Tompson*.

(Note that in these cases we can be sure of the direction of change, since *Tomkin* is clearly derived from ‘kin of Tom’ and *Tomson* from ‘son of Tom’. In other cases, it may be less clear. *Sempstress* is derived from *seam+stress*; *Hampshire* probably from *Ham+shire*.)

10. a) It seems that /d/ remains [d] at the start of a word but becomes [ð] elsewhere. Formally:

$$/d/ \rightarrow [\text{ð}] : \text{not-start-of-word } _$$

The odd thing about this rule is its lack of motivation. Most of the rules found in earlier exercises change properties of phones (e.g. their position of articulation) rather than just arbitrarily changing one phone into another. Why would an alveolar stop change to a dental fricative?

- b) If however Spanish has the phoneme [d̪], then a more plausible rule can be written:

$$\{\text{stop, dental, voiced}\} \rightarrow \{\text{fricative, dental, voiced}\} : \text{not-start-of-word } _$$

- c) To generalize, we can re-write the rule as:

$$\{\text{stop, Posn, voiced}\} \rightarrow \{\text{fricative, Posn, voiced}\} : \text{not-start-of-word } _$$

This suggests that the /g/ in *paga* should become a voiced velar fricative ([ɣ] in IPA) and the /b/ in *dividir* a voiced bilabial fricative ([β] in IPA). Neither phone occurs in English; [ɣ] is like a *g* in the same throat position as the *ch* in the Scottish *loch*; for [β], try saying *v* with both your lips rather than with your upper lip and lower teeth. Try to find a Spanish speaker from central/northern Spain to confirm or refute this analysis! (Those who complain about English spelling should note that in the Spanish word *viva*, neither *v* is pronounced as [v], since its ‘Standard Castilian’ pronunciation is [biBa]!)

Morphology

1. a) Words showing inflectional morphology are:

Word	Lexeme	Affix
<i>been</i>	<i>be</i>	<i>en</i>
<i>being</i>	<i>be</i>	<i>ing</i>
<i>components</i>	<i>component</i>	<i>s</i>
<i>elements</i>	<i>element</i>	<i>s</i>
<i>includes</i>	<i>include</i>	<i>s</i>
<i>languages</i>	<i>language</i>	<i>s</i>
<i>ordering</i>	<i>order</i>	<i>ing</i>
<i>others</i>	<i>other</i>	<i>s</i>
<i>retained</i>	<i>retain</i>	<i>ed</i>
<i>rules</i>	<i>rule</i>	<i>s</i>
<i>sentences</i>	<i>sentence</i>	<i>s</i>
<i>words</i>	<i>word</i>	<i>s</i>

The first example is arguable. The affix *en* is used with only a few verbs (e.g. *eat* – *eaten*, *beat* – *beaten*) to construct the form used in a sentence containing *I have* – (e.g. *I have been happy*, *I have eaten too much*). Most verbs in Modern English either have the affix *ed* in this context (e.g. *I have finished*, *I have included this example*) or have an irregular form (e.g. *I have thought about it*, *I have drunk all the wine*). Hence *been* might equally have been treated as an irregular form.

- b) Derivational morphology is a fuzzy area. Here are some possibilities:

Word	Derivation
<i>against</i>	<i>again</i> + <i>st</i>
<i>another</i>	<i>an</i> + <i>other</i>
<i>Indo-European</i>	<i>Indo</i> + <i>Europe</i> + <i>an</i>
<i>morphology</i>	<i>morph</i> + <i>ology</i>
<i>notable</i>	<i>note</i> + <i>able</i>
<i>Persian</i>	<i>Persia</i> + <i>an</i>
<i>Russian</i>	<i>Russia</i> + <i>an</i>
<i>within</i>	<i>with</i> + <i>in</i>

Some of these are arguable. In Modern English, *again* and *again+st* don't seem to be related. *Another* and *within* could be argued to be combinations of two lexemes, rather than derivations from one core lexeme. Others could be broken down further, e.g. *Indo-* and *-ology* contain the infix *-o-* which is often used to combine words, particularly names of countries (e.g. *French* + *German* = *Franco-German*).

One the other hand, *English* looks as though it might be derived from *Engl* + *ish* (*ish* is a fairly common affix, e.g. *greenish*, *largish*, etc.). Historically this is true, the root lexeme being *Engle* or *Angle* (as in e.g. *East Anglia*). *Angl(e)* does generate a number of words, e.g. the prefix *Anglo-*, *Anglia*, *Anglian*, *Anglican*. *Engl(e)* on the other hand is only found again in *England*, which looks as if it is *Eng* + *land*, rather than *Engl* + *and*. (Historically it was *Englaland* or *Anglaland* – the land of the Engles or Angles, the *a* having disappeared to make pronunciation easier.)

All this goes to show that trying to write rules for computer processing of derivational morphology will be difficult, if not impossible. On the other hand, human abilities in this area are important. For example, just before I first wrote this I heard, for the first time so far as I can recall, the word *novelization*, meaning (presumably) to make a novel out of something (the context was *the novelization of 'The Archers'*).

2. The table below shows the written morphemes with their correspondences. (' \emptyset ' = nothing.)

<i>skil+os</i>	<i>dog+\emptyset</i>	<i>gat+a</i>	<i>cat+\emptyset</i>
<i>skil+o</i>	<i>dog+\emptyset</i>	<i>gat+as</i>	<i>cat+'s</i>
<i>skil+ou</i>	<i>dog+'s</i>	<i>gat+es</i>	<i>cat+s</i>
<i>skil+oi</i>	<i>dog+s</i>	<i>gat+on</i>	<i>cat+s'</i>
<i>skil+ous</i>	<i>dog+s</i>		
<i>skil+on</i>	<i>dog+s'</i>		

[On the evidence presented here, it would also be correct to treat the root of the Greek for dog as *skilo*, so that e.g. *skilos* = *skilo+s*. In the Greek alphabet however, *skilos* is written $\sigma\kappa\acute{\upsilon}\lambda\omicron\varsigma$, whereas *skilon* is written $\sigma\kappa\acute{\upsilon}\lambda\omicron\nu$, showing that the root is $\sigma\kappa\acute{\upsilon}\lambda$ which I have transcribed as *skil*.]

3. Note that the possessive formed in this way is more common with nouns representing animate objects. *The girl's hair* is more natural than *the cinema's roof* which would, in writing at least, more often be expressed as *the roof of the cinema*. Hence my examples below deliberately use animate objects.

- a) 1. To form the possessive singular, add 's to the singular (base) form of the noun.
Examples:

dog's cat's fox's hippopotamus's¹ mouse's man's sheep's

2. To form the possessive plural, first form the plural of the noun.
– If the plural ends in *s* then add an apostrophe only (i.e. ').
– If the plural does not end in *s* then add 's.

Examples:

dogs' cats' foxes' hippopotamuses' mice's men's sheep's

- b) 1. To form the possessive singular, follow EXACTLY the same rules as for the regular plural. That is, add /z/, then insert /ɪ/ if the /z/ is preceded by an alveolar or palatal fricative or affricative, then de-voice the /z/ to /s/ if it is preceded by a voiceless phone. Examples:

[dɒgz] [cæts] [maʊsɪz] [mænz] [ʃɪps]

2. To form the possessive plural, first form the plural of the noun.
– If the plural was formed by the regular 'add /z/' rule then do nothing more.
– If the plural is irregular, then apply the regular 'add /z/' rule.

Examples:

[dɒgz] [cæts] [maɪsɪz] [mɛnz] [ʃɪps]

The key point seems to be that the 'add /z/' rule is applied only ONCE. Thus in speech, the plural, possessive singular and possessive plural are regularly formed in exactly the same way, i.e. by the application of the 'add /z/' rule. If the plural is formed by a different rule, then the 'add /z/' rule can be used to form the possessive plural.

See also the answer to Exercise 6.

4. We can either write 'add *ed*' rules, or assume the *ed* is added and then write 'change' rules. I'll do the latter. Let V stand for any vowel letter, C any consonant letter (y here is a consonant), L any letter. Three rules are:

$L e ed \rightarrow L ed$

$C y ed \rightarrow C i ed$

$C_1 V C_2 ed \rightarrow C_1 V C_2 C_2 ed$ ($C_2 \neq w, x$ or y)

If none of these rules apply, the result of the addition of *ed* remains unchanged.

¹ An alternative rule omits the *s* when the singular ends in *s*, particularly when preceded by a vowel. Thus *the method of Socrates* will be written as *Socrates' method* with the pronunciation [sʌkr'tɪz], perhaps because [sʌkr'tɪzɪz] might suggest that 'add /z/' rule has been applied twice which is not allowed in English.

A possible algorithm is:

```

add_ed(Word):
  Word → Word + "ed";
  FOR each rule in the rule set LOOP
    IF the last letters of Word match the lhs of rule THEN
      use the rhs of rule to change Word;
      RETURN Word;
    END IF
  END LOOP
  RETURN Word;
END add_ed

```

The complete algorithm to find the past of a verb will be something like:

```

past(Verb):
  IF Verb is in the lexicon with the irregular past Past THEN
    RETURN Past;
  ELSE
    RETURN add_ed(Verb);
  END IF
END past

```

5. Firstly, we will need an algorithm to remove *ed*, e.g.:

```

rem_ed(Word):
  FOR each rule in the rule set LOOP
    IF the last letters of Word match the rhs of rule THEN
      use the rhs of rule to change Word;
      RETURN Word - "ed";
    END IF
  END LOOP
  RETURN Word;
END rem_ed

```

The complete algorithm can take one of two forms. One possibility is:

```

base(Past):
  IF Past is in the lexicon as the irregular past of Base THEN
    RETURN Base;
  ELSE
    RETURN rem_ed(Past);
  END IF
END base

```

Note that `base("caught")` will yield "catch" using this algorithm – since "caught" will not be in the lexicon, `rem_ed` will be applied. If it is desired to prevent this, then the algorithm must be changed to:

```

base(Past):
  IF Past is in the lexicon as the irregular past of Base THEN
    RETURN Base;
  ELSE
    Base → rem_ed(Past);
    IF Base is not in the lexicon with an irregular past THEN
      RETURN Base;
    ELSE
      fail;
    END IF
  END IF
END base

```

That is we must check that the derived base form does not have an irregular form in the lexicon. Since this algorithm requires two lexicon look-ups, efficiency will be important (e.g. the use of a hash table).

6. The rules I seem to apply in fast speech for plurals are those in S1 to S3, with one extension. For unvoiced labio-dental and dental fricatives, it depends on the length of previous vowel.

The consistently 'long' vowels in SEE are /a/, /u/, /ɔ/ and all the diphthongs; /i/ varies but is usually 'long'. The rest are 'short'. Note that length is never the only difference between the vowel phonemes in English.

If the previous vowel is short, S2 applies. However, if it is long, S3 applies but the fricative is voiced as well. So the table below shows my answer – however, there are exceptions (e.g. *laughs* is [lafɪs] not [lavz]). Your answers could be different.

Last phone(s)	Example Singular – Plural	My pronunciations
[p]	<i>cup – cups</i>	[kʌp] – [kʌps]
[b]	<i>cab – cabs</i>	[kæb] – [kæbz]
[t]	<i>cat – cats</i>	[kæt] – [kæts]
[d]	<i>lad – lads</i>	[læd] – [lædz]
[k]	<i>duck – ducks</i>	[dʌk] – [dʌks]
[g]	<i>dog – dogs</i>	[dɒg] – [dɒgz]
[m]	<i>brim – brims</i>	[brɪm] – [brɪmz]
[n]	<i>run – runs</i>	[rʌn] – [rʌnz]
[ŋ]	<i>rung – rungs</i>	[rʌŋ] – [rʌŋz]
short-vowel [f]	<i>cough – coughs</i>	[kɒf] – [kɒfs]
long-vowel [f]	<i>loaf – loaves</i>	[ləʊf] – [ləʊvz]
[v]	<i>groove – grooves</i>	[gru:v] – [gru:vz]
short-vowel [θ]	<i>moth – moths</i>	[mɒθ] – [mɒθs]
long-vowel [θ]	<i>mouth – mouths</i>	[maʊθ] – [maʊðz]
[ð]	<i>scythe – scythes</i>	[saɪð] – [saɪðz]
[s]	<i>bus – buses</i>	[bʌs] – [bʌsɪz]
[z]	<i>buzz – buzzes</i>	[bʌz] – [bʌzɪz]
[ʃ]	<i>bush – bushes</i>	[bʊʃ] – [bʊʃɪz]
[ʒ]	<i>mirage – mirages</i>	[mɪrɑʒ] – [mɪrɑʒɪz]
[tʃ]	<i>latch – latches</i>	[lætʃ] – [lætʃɪz]
[dʒ]	<i>judge – judges</i>	[dʒʌdʒ] – [dʒʌdʒɪz]

If we assume that the plural is always formed by adding /z/, then in my dialect, the following phonological rules apply.

P1 /z/ → [ɪz] : {fricative | affricative, alveolar | palatal} _ end of word

P2 /z/ → [s] : {stop, unvoiced} _ end of word

P3 /z/ → [s] : {short-vowel} {fricative, labio-dental | dental, unvoiced} _ end of word

P4 {fricative, labio-dental | dental, unvoiced} →
 {fricative, labio-dental | dental, voiced} : {long-vowel} _ /z/ end of word

The key feature seems to be that the last two phones must agree in voicing: either both unvoiced or both voiced. Note how complicated even such a simple action as pronouncing the plural of a noun turns out to be. However, although complex, it appears to be rule-determined and thus programmable.

Additional Notes to Exercise 3b

Originally the *gh* in words like *cough* or *laugh* represented a sound like the German *ch* in *nach* or *ich*. At the end of words, it was sometimes replaced in Middle and Modern English by [f]. Interestingly, I seem to pronounce the plurals of such words with [fs] regardless of the length of the preceding vowel, e.g. *laughs* is [lafɪs] whereas *half* [hʌf] has the plural [hʌvz]. Spelling seems to be the determining factor here.

Some authors claim that in spoken English the rules for forming the possessive singular and the plural are the same. However, at least in my idiolect/dialect, this is not correct, since I pronounce *wife's* or *path's* differently from the plurals *wives* and *paths*. I don't apply Rules P3 and P4 above to possessive singulars. English spelling shows my rules for *f/v*, but cannot for [θ]/[ð]. However, I don't think I am entirely consistent. At the time of writing I haven't found an account of this elsewhere so I would be very interested to know whether other SEE speakers agree or disagree. Note that the situation is different in Northern English English.

Syntax

1.
 - a) *Some*_{DET} *people*_N *like*_V *cats*_N.
 - b) *Europeans*_N *peopled*_V *America*_N.
 - c) *Careful*_A *owners*_N *wash*_V *their*_{DET} *cars*_N.
 - d) *Down*_N *fills*_V *the*_{DET} *best*_A *duvets*_N.
 - e) *She*_{PRN} *might*_{AUX} *drive*_V *down*_P *my*_{DET} *street*_N.
 - f) *The*_{DET} *man*_N *with*_P *a*_{DET} *wooden*_A *leg*_N *ate*_V *my*_{DET} *hamburger*_N.
 - g) *No-one*_{PRN} *saw*_V *her*_{PRN}.
 - h) *You*_{PRN} *should*_{AUX} *put*_V *paint*_N *on*_P *the*_{DET} *sound*_A *wood*_N.
 - i) *I*_{PRN} *heard*_V *a*_{DET} *wooden*_A *sound*_N.
 - j) *The*_{DET} *bell*_N *sounds*_V *for*_P *tea*_N.
 - k) *I*_{PRN} *have*_{AUX} *painted*_V *the*_{DET} *outside*_N *of*_P *my*_{DET} *house*_N.
 - l) *I*_{PRN} *put*_V *the*_{DET} *tub*_N *of*_P *red*_A *geraniums*_N *outside*_P *my*_{DET} *house*_N.

2.
 - a) *(Some people)*_{NP} *(like cats)*_{NP}_{VP}
 - b) *Europeans*_{NP} *(peopled America)*_{NP}_{VP}
 - c) *(Careful owners)*_{NP} *(wash (their cars))*_{NP}_{VP}
 - d) *Down*_{NP} *(fills (the best duvets))*_{NP}_{VP}
 - e) *She*_{NP} *(might drive (down (my street)))*_{NP}_{PP}_{VP}
 - f) *(The man (with (a wooden leg)))*_{NP}_{PP}_{NP} *(ate (my hamburger))*_{NP}_{VP}
 - g) *No-one*_{NP} *(saw her)*_{NP}_{VP}
 - h) *You*_{NP} *(should put paint (on the sound wood))*_{NP}_{VP}
 - i) *I*_{NP} *(heard (a wooden sound))*_{NP}_{VP}
 - j) *(The bell)*_{NP} *(sounds (for tea))*_{PP}_{VP}
 - k) *I*_{NP} *(have painted (the outside (of (my house)))*_{NP}_{PP}_{NP}_{VP}
 - l) *I*_{NP} *(put (the tub (of (red geraniums)))*_{NP}_{PP}_{NP} *(outside (my house))*_{NP}_{PP}_{VP}

3.
 - a) Here's one grammar:
 - S → NP VP
 - NP → noun
 - VP → verb
 - VP → verb NP
 - noun → {any word in the lexicon as a noun}
 - verb → {any word in the lexicon as a verb}

An simpler grammar replaces the first four productions above by:

 - S → noun verb
 - S → noun verb noun

In either case, the lexicon could be:

 - anglers* : noun
 - eat* : verb
 - fish* : noun
 - fish* : verb
 - otters* : noun
 - swim* : verb

Either grammar generates sentences like *anglers swim fish*, in which the verb has the wrong complement (i.e. NP instead of nothing). This is a syntactic error. A sentence like *otters eat anglers* might be said to be semantically anomalous (although there could be circumstances in which it was true..).

- b) With the first grammar we re-write the VP productions:

```
VP → verb
VP → aux verb
VP → verb NP
VP → aux verb NP
```

A slightly neater alternative (but linguistically more suspect) is:

```
VP → verb_group
VP → verb_group NP

verb_group → verb
verb_group → aux verb
```

In either case we need to add:

```
aux → {any word in the lexicon as an auxiliary}
```

```
can : aux
do  : aux
.....
```

- c) This gets more complex! There must be hundreds of ways of writing the required grammar. One approach is to start by expanding s into two different kinds of sentence:

```
S → S_dec           (dec for declarative)
S → S_int           (int for interrogative)

S_dec → NP SVP
S_dec → NP aux SVP
S_dec → NP aux not SVP

S_int → aux NP SVP
S_int → aux not NP SVP
S_int → aux NP not SVP

SVP → verb
SVP → verb NP
```

Finally we need a rule which says that the sequence *aux not* can be replaced by *auxn't*. Note that this is a context sensitive rule. This grammar allows sentences like *Do otters not eat fish?* and *Do not otters eat fish?* which are rare in modern English. Whatever grammar you wrote, make sure it can't generate two occurrences of *not*.

An alternative approach is 'transformational'. We write only the grammar for *S_dec*. Next we have the optional *aux not* → *auxn't* rule. Finally we have a transformational rule that says to form an interrogative sentence, the first NP is swapped with the next element if this is *aux* or *auxn't*. This grammar allows *Do otters not eat fish?* (non-application of the optional *n't* rule followed by the swap rule) but not *Do not otters eat fish?* (because only one element can be swapped with the first NP).

4. Here's one answer. The grammar could be:

```
S → NP(N) VP(N)

VP(N) → verb(N) PP

NP(N) → det(N) noun(N)
NP(p) → noun(p)
```

PP → prep NP(_)

det(N) → {any word in the lexicon as a det of number N}

noun(N) → {noun, N}

verb(N) → {verb, N}

prep → {prep}

A lexicon for the words in the question might be:

```
% DETERMINERS - WORD : det,NUMBER
```

```
a : det,s
```

```
the : det,_
```

```
% NOUNS - WORD : noun,NUMBER
```

```
dog : noun,s
```

```
dogs : noun,p
```

```
floor : noun,s
```

```
floors : noun,p
```

```
basket : noun,s
```

```
baskets : noun,p
```

```
% VERBS - WORD : verb,NUMBER
```

```
sits : verb,s
```

```
sit : verb,p
```

```
is : verb,s
```

```
are : verb,p
```

```
% PREPOSITIONS - WORD : prep
```

```
in : prep
```

```
on : prep
```

5. a) Using brackets rather than drawing a tree:

$((sensei_{noun} wa_{part})_{PartP} ((honsha_{noun} ni_{part})_{PartP} iru_{verb})_{VP})_S$ or
 $S(PartP(noun(sensei),part(wa)),VP(PartP(noun(honsha),part(ni)),verb(iru)))$

b) *honsha ni iru* is roughly equivalent to ‘he’s in the office’, where *he* isn’t stressed. Implicit subjects may be omitted in Japanese.

honsha wa sensei ni iru is syntactically correct according to this grammar but would mean ‘the office is in the teacher’.

6. d) Add an extra `verbComps` production:

```
verbComps(s) → that S | S
```

Then add the new verbs to the lexicon:

```
% VERBS - WORD : verb,Person,Number,ValidComplementType
```

```
know : verb,1,s,s
```

```
know : verb,2,s,s
```

```
knows : verb,3,s,s
```

```
know : verb,_,p,s
```

```
knew : verb,_,_,s
```

A better approach of course would be to put the ‘base’ verb *know* in the lexicon as having the valid complement type `s` and then use morphological processing to handle the person and number forms.

The grammar is ‘automatically’ recursive, allowing sentences such as *They knew that I knew that she thought that she gave the bottle to me.*

7. It seems to me to work for the following.

- `s`, since we can have sentences such as *She gave the bottle to us and we gave the bottle to them.* The production is:

```
s → s and s
```

- NP, since we can have sentences such as *The man and the woman gave...* or *The man and I gave...* So if NP has agreement in person, number and case, we can have the production:

$$\text{NP}(?, p, C) \rightarrow \text{NP}(P_1, N_1, C) \text{ and } \text{NP}(P_2, N_2, C)$$

The co-ordinated NP is clearly plural. The only problem relates to its person. One rule is that the final person is numerically the lower of the persons involved. Thus *you and I* is equivalent to *we*, *you and he* to *you*.

- prn (but this production is unnecessary, since it will be generated by co-ordinating NPs).
- As noted in the question it works for noun, giving *the old man and woman*.
- It works for adj, as in e.g. *the old and weary man*. (But more usual is co-ordination of only the last in a list of adjs, e.g. *the old, weary and unhappy man*.)

$$\text{adj} \rightarrow \text{adj and adj} \mid \{\text{any word in the lexicon as an adj}\}$$
- It works for VP, as in e.g. *The man saw the cat and gave it a bone*.

$$\text{VP}(P, N) \rightarrow \text{VP}(P, N) \text{ and } \text{VP}(P, N)$$
- It works for verb, as in e.g. *The man saw and heard the cat*.

$$\text{verb}(P, N) \rightarrow \text{verb}(P, N) \text{ and } \text{verb}(P, N)$$
- It works for VerbComps, as in e.g. *The man gave a bone to the cat and a biscuit to the dog*.

$$\text{verbComps}(CType) \rightarrow \text{verbComps}(CType) \text{ and } \text{verbComps}(CType)$$
- It works for PP, as in e.g. *The man gave a bone to the cat and to the dog*.

$$\text{PP} \rightarrow \text{PP and PP}$$
- It works for prep, as in e.g. *I looked on and under the table*.

The only case where it doesn't seem to work is det. We can't say *this and that man* for example. (What about *Both these and those cats are yours*?) Co-ordinated SNPs are uncommon, but possible in restricted circumstances. We don't usually say *the old man and young woman* but *the old man and the young woman* (co-ordinating NPs rather than SNPs). But I can say *the girl with the red blouse and long skirt*. Some sense of 'pairing' seems to be needed to allow co-ordination of SNPs.

NOTE: in practice, serious problems can arise if productions like these are included in a grammar. 'Left recursive' productions, e.g.

$$a \rightarrow a p$$

can result in infinite loops if used with a top-down, left-right parsing algorithm.

8. a) (ii), (iii) and (viii) are invalid. The rest are valid.
9. a) For my grammar, we might obtain the following. Note that NP1 and NP2 do not appear in the tree: their sole purpose is to get the order right. I have made sure that the VP tree is always in the order verb followed by verb complement. Just for variety, I've added the tree variable at the end of any other arguments.

$$\text{F_S}(s(T_1, T_2)) \rightarrow \text{F_NP}(P, N, \text{nom}, T_1) \text{ F_VP}(P, N, T_2)$$

$$\text{F_NP}(P, N, C, T) \rightarrow \text{F_NP1}(P, N, C, T)$$

$$\text{F_NP}(3, N, _, T) \rightarrow \text{F_NP2}(N, T)$$

$$\text{F_NP1}(P, N, C, \text{np}(T)) \rightarrow \text{f_prn}(P, N, C, T)$$

$$\text{F_NP2}(N, \text{np}(T_1, T_2)) \rightarrow \text{f_det}(N, T_1) \text{ f_noun}(N, T_2)$$

$$\text{F_VP}(P, N, \text{vp}(T_1, T_2)) \rightarrow \text{F_NP1}(_, _, \text{acc}, T_2) \text{ f_verb}(P, N, T_1)$$

$$\text{F_VP}(P, N, \text{vp}(T_1, T_2)) \rightarrow \text{f_verb}(P, N, T_1) \text{ f_NP2}(_, T_2)$$

```

f_prn(P,N,C,prn(Word)) → {F_Word : prn,P,N,C,_}
f_det(N,det(Word))    → {F_Word : det,N,_}
f_noun(N,noun(Word))  → {F_Word : noun,N,_}
f_verb(P,N,verb(Word)) → {F_Word : verb,P,N,_}

```

- b) *Le chat me voit* will produce $s(np(det(le), noun(chat)), vp(verb(voit), np(prn(me))))$. Note that I have arranged that the grammar always produces a tree with the complement after the verb, regardless of the order in the input sentence. So simply replacing the French words with English ones to give $s(np(det(the), noun(cat)), vp(verb(see), np(prn(me))))$ and then putting the tree back through an appropriate English grammar should succeed.

When trees produced by the grammars for each language are identical, translation needs only to translate the words. Note that this needs to be done for ALL inflectional variants. Clearly it is better to store base lexemes in the tree and use morphological rules to generate actual words. If on the other hand, the grammar stored the French (sub)tree for pronoun objects in a different order, then a ‘transfer’ rule would be needed. For example, the French tree:

```
vp(np(prn(FrPronoun)), verb(FrVerb))
```

would need to be mapped into the English tree:

```
vp(verb(EnVerb), np(prn(EnPronoun)))
```

Meaning

1. Here are four possible rules:

-COMMON \Rightarrow [-ABSTRACT]
 +ANIMATE \Rightarrow [-ABSTRACT]
 +MALE \Rightarrow [+ANIMATE, -FEMALE, -ABSTRACT]
 +FEMALE \Rightarrow [+ANIMATE, -MALE, -ABSTRACT]

2. One possibility, given just these examples, is:

featureList1 = [+PLURALFORM | -COMMON, +RIVER]
 featureList2 = [-PLURALFORM, -COMMON, -RIVER]

PLURALFORM is needed rather than grammatical number, because *United States* is singular in that it requires a singular verb to follow (since we say *The United States is a very large country*, not *The United States are a very large country*).

Jane : noun, [-PLURALFORM, -COMMON, -RIVER]
Everest : noun, [-PLURALFORM, -COMMON, -RIVER]
Himalayas : noun, [+PLURALFORM, -COMMON, -RIVER]
Thames : noun, [-PLURALFORM, -COMMON, +RIVER]
London : noun, [-PLURALFORM, -COMMON, -RIVER]
England : noun, [-PLURALFORM, -COMMON, -RIVER]
United States : noun, [+PLURALFORM, -COMMON, -RIVER]

Note that it's important to have the 'or' in featureList1. The lexicon lookup algorithm will have to ensure correct feature matching.

[Seas and oceans also select *the* as in, for example, *The Atlantic is an ocean* or *The Mediterranean is a sea*. It is tempting to think that the relevant semantic feature is +WATER rather than +RIVER but lakes, fjords, etc. do not require *the*. Thus *Loch Lomond is a Scottish lake* or *Hardanger is one of the most beautiful Norwegian fjords*. Why ARE NLS so complicated?!

3. These quantifiers fall into two groups:

- a) Selecting +PLURAL OR -COUNT. (Note that +PLURAL \Rightarrow +COUNT, -COUNT \Rightarrow -PLURAL.)
All and *most*.
 b) Selecting -COUNT.
Much and *less*.

Some belongs to group (a). *Little* belongs to (b). There are at least three further groups:

- c) No selection, i.e. allowing singular or plural, \pm COUNT.
No, *any*.
 d) Selecting -PLURAL and +COUNT.
Each, *every*, *one*.
 e) Selecting +PLURAL (hence +COUNT).
Both, *several*, *few*, *many*, and all numbers from *two* upwards.

[If we go on to consider combining a quantifier with *the*, the situation becomes even more complicated. Some quantifiers can be preceded by *the* (e.g. *The few boys who came left early*), others can be followed by *the* (e.g. *All the boys came*), others cannot combine with *the* (e.g. *no*). These distinctions cut across the groups set out above.]

4. One possible answer is to have lexicon entries as follows:

colourless : [-ABSTRACT, -COLOUR]
green : [-ABSTRACT, +COLOUR]
idea : [+ABSTRACT]
sleep : [-ACTIVITY], requires subject NP: [-ABSTRACT]
furiously : [+ACTIVITY]

Colourless green is then unacceptable because -COLOUR clashes with +COLOUR. *Colourless ideas* and *green ideas* are unacceptable because -ABSTRACT clashes with +ABSTRACT. *Ideas sleep* is unacceptable because the subject NP must be -ABSTRACT but is actually +ABSTRACT. *Sleep furiously* is unacceptable because *sleep* has the semantic feature [-ACTIVITY] but *furiously* is +ACTIVITY.

The danger with this approach is of inventing arbitrary semantic features which are then used with only a few words, thereby introducing redundancy and explaining very little.

5. a) *eat* : complements = NP/Agent:[+ANIMATE] + {NP/Patient} + {NP/Instrument}

Children eat sweets. Agent = *children*; Patient = *sweets*.

Sweets are eaten by children. Patient = *sweets*; Agent = *children*.

Sweets are eaten. Patient = *sweets*; Agent omitted.

**Children are eaten by sweets.* Agent must be animate.

Sweetcorn is eaten with the fingers. Patient = *sweetcorn*; Agent omitted; Instrument = *the fingers*.

The children are eating. Agent = *the children*.

**Sweets are eating.* Agent must be animate.

It seems desirable to restrict the Patient to avoid e.g. *Books are eaten* or *Children are eaten*, but I'm not convinced by any attempt to do so which I've seen. The Patient can be either animate (since e.g. I have eaten live shellfish) or inanimate (e.g. sweets). You could try something like +EDIBLE, but could every lexicon entry be consistently labelled? Also there are strong cultural and religious differences affecting the edibility of foods.

- b) If we give *open* the same verb complements as *break*, we have:

break : complement = NP/Agent:[+ANIMATE] + NP/Patient:[-ANIMATE] +
{PP/Instrument:[-ANIMATE]} |

NP/Force:[-ANIMATE] + NP/Patient:[-ANIMATE] |

NP/Patient:[-ANIMATE]

The first complement explains sentences like:

The boy opened the door. Agent = *the boy*; Patient = *the door*.

**The door opened the boy.* Patient must be -ANIMATE.

The door was opened by the boy. Patient = *the door*; Agent = *the boy*.

The door was opened. Patient = *the door*; Agent missing.

**The boy was opened.* Patient must be +ANIMATE.

The boy opened the door with the key. Agent = *the boy*; Patient = *the door*; Instrument = *the key*.

The door was opened with a key by the boy. Patient = *the door*; Instrument = *the key*; Agent = *the boy*.

The door was opened with a key. Patient = *the door*; Instrument = *the key*; Agent missing.

The second complement accounts for sentences like:

The wind opened the door. Force = *the wind*; Patient = *the door*.

The door was opened by the wind. Force = *the wind*; Patient = *the door*.

The key opened the door. Force = *the key*; Patient = *the door*.

?*The door was opened by the key.* Force = *the key*; Patient = *the door*.

I'm not entirely convinced by this analysis. Some English sentences with no Agent show clear differences in the use of *with* and *by* followed by an inanimate NP. Thus there are sentences where only one of these prepositions is acceptable. Only when *by* is possible can the following NP be made the subject:

*The window was broken (by / *with) the wind.* (*the wind* = Force, not Instrument)

The wind broke the window. (Passive allowed because *the wind* = Force)

*She was tied up (*by / with) the rope.* (*the rope* = Instrument, not Force)

**The rope tied her up.* (Passive not allowed because *the rope* = Instrument not Force).

This suggests a clear difference between Force and Instrument. But I can't find any examples where *was opened by the key* and *was opened with the key* have clearly different meanings, and the former sounds odd to me. Perhaps we should allow *open* to have Instrument as well as Force as its subject in an active sentence. This suggests a slight difference between the two verbs.

The final complement generates:

The door opened. Patient = *the door*.

**The boy opened.* Patient must be -ANIMATE.

[To rule out *The key was opened with the door* we would need further semantic properties. Note again the subtle difference between *the door opened* and *the door was opened*. In the former, there is no Agent – the implication is that the door ‘just opened’; in the latter, there is an Agent, but it is not given in the sentence – something opened the door, but we aren't told what.]

c) *give* : complements =

NP/Agent: [+ANIMATE] + NP/Patient + PP/Recipient: [+ANIMATE] |
NP/Agent: [+ANIMATE] + NP/Recipient: [+ANIMATE] + NP/Patient

The first complement explains sentences such as:

Jo gave the cup to me. Agent = *Jo*; Patient = *the cup*; Recipient = *me*.

**Jo gave to me.* Patient missing.²

It was given to me by Jo. Agent = *Jo*; Patient = *it*; Recipient = *me*.

It was given to me. Patient = *it*; Recipient = *me*; Agent missing.

It was given to me by Jo. Agent = *Jo*; Patient = *it*; Recipient = *me*.

**I was given by the ball.* Agent not animate; Patient missing.

The second complement is interesting because it allows passivization with the Recipient becoming the grammatical subject (usually the Patient becomes the grammatical subject):

Jo gave me the ball. Agent = *Jo*; Recipient = *me*; Patient = *it*.

I was given the ball by Jo. *I* = Recipient; *the ball* = Patient; Agent *Jo*.

I was given the ball. *I* = Recipient; *the ball* = Patient; Agent missing.

**The ball was given me by Jo.* Not a valid passive complement.³

[But I think I would say *That was given me by my wife* when speaking informally, although I might ‘correct’ it in writing. Language is difficult to describe by rigid rules!]

d) Suppose that we describe *throw* as follows.

throw : complements =

NP/Agent: [+ANIMATE] + NP/Patient + {PP/Source} + {PP/Goal} |
NP/Agent: [+ANIMATE] + NP/Patient + PP/Recipient: [+ANIMATE] |
NP/Agent: [+ANIMATE] + NP/Recipient : [+ANIMATE] + NP/Patient³

The first complement corresponds to that for *push* (i.e. PROPEL in CD theory), the last two to those for *give* (i.e. PTRANS in CD theory).

Joe threw the ball. Agent = *Joe*; Patient = *the ball*.

**Joe threw.* Patient missing.

**The garden threw the ball.* Agent not animate.

The ball was thrown by Joe. Agent = *Joe*; Patient = *the ball*.

The ball was thrown. Patient = *the ball*; Agent missing.

Joe threw it into the garden. Agent = *Joe*; Patient = *it*; Goal = (*into*) *the garden*.

² But there are cases when missing Patient is allowed, e.g. *I never give to beggars*. This seems possible where what is being given (money) can be taken for granted.

³ Note that this complement allows passivization. It is this which distinguishes Goal from Recipient, since the Goal cannot be made the subject: **The garden was thrown the ball* is semantically anomalous and anyway is not the passive of *X threw the ball into the garden*.

The ball was thrown out of the house into the garden. Patient = Joe; Source = (*out of*) the house; Goal = (*into*) the garden; Agent missing.

Joe threw the ball to me. Agent = Joe; Patient = the ball; Recipient = me.

**Joe threw to me.* Patient missing.

It was thrown to me by Joe. Agent = Joe; Patient = it; Recipient = me.

Joe threw me the ball. Agent = Joe; Recipient = me; Patient = it.

I was thrown the ball. I = Recipient; the ball = Patient.

**I was thrown by the ball.* Agent not animate.

It seems to work, suggesting that *throw* does have both PROPEL and PTRANS meanings.

This analysis can be criticized. If I throw something to you, you don't necessarily have it, whereas you do if I give it to you, suggesting that *throw* doesn't exactly have a Recipient. *Push* sometimes seems to allow a Recipient, e.g. *He pushed the box to me (so I could look at it)*. However, note that *throw* allows a Recipient noun phrase, but *push* doesn't (at least it seems very odd to me):

Joe threw me the box.

**Joe pushed me the box.*

The moral, yet again, is that language is hard to capture by rules!

- e) At CAN introduce a Location, which is then questioned by *Where?*, as in:

I put the ball at the bottom of the tree. Where did you put the ball? At the bottom of the tree.

However it only seems to do this in certain fixed phrases (e.g. *at the top of*, *at the bottom of*), whereas in the sentences given here *at* can be followed by any semantically relevant noun phrase. Furthermore, *Where?* is not the right question for this use of *at*:

*I drove my car at the policeman. Where did you drive your car? *At the policeman.*

I drove my car at the policeman. Who⁴ did you drive your car at? The policeman.

Another possibility seems to be Goal. Against this, this kind of PP doesn't combine well with Source:

?*I drove my car out of the garage at the policeman.*

Furthermore, Goal is usually questioned by *Where (to)?*:

I drove my car into the garage. Where did you drive your car? Into the garage.

*I drove my car at the garage. Where did you drive your car? *At the garage.*

I drove my car at the garage. What did you drive your car at? The garage.

I conclude that we have a new θ -role, which we might call Target: it's the object at which the Patient of a PROPEL verb can be aimed; the intention of the Agent is to touch/hit the Target with the Patient. (Again the need for this extra θ -role raises the question of whether there is a finite set of such roles.) *Throw* then needs an extra complement:

NP/Agent: [+ANIMATE] + NP/Patient + PP (*at*)/Target

giving analyses like:

Joe threw the ball at the wall. Agent = Joe; Patient = the ball; Target = the wall.

Joe threw it at me. Agent = Joe; Patient = it; Target = me.

It was thrown at the wall. Agent missing; Patient = it; Target = the wall.

⁴ And yes, I do know that pedants will argue for *whom* here. But this is just pedantry. Some English speakers speaking formally do regularly use a rule which says that when *who* immediately follows a preposition to form a PP it becomes accusative (as would a pronoun). Thus *To whom did you give you the revolver?* But even these speakers will regularly say *Who did you give the revolver to?*

The difference between Goal and Target is crucial in these sentences:

He drove his car towards the policeman. The policeman = Goal

He drove his car at the policeman. The policeman = Target