Natural Language Processing & Applications

Morphology

1 Introduction

As used in this module, one use of the word 'grammar' is to refer to systematic patterns both within words and within sentences. The grammar of a language thus includes both the **morphology** of words and the **syntax** of sentences. Languages tend to 'trade off' these components against one another. Some Indo-European languages (Russian being a notable example) have retained the complex morphology which is thought to have been a feature of the original Indo-European language. Others, such as English and Persian (Farsi), have less complex morphology but more rigid syntax (rules ordering elements within sentences). Thus the word order in the English sentence:

The dog saw the cat.

is absolutely fixed. Swapping *the dog* for *the cat* produces an equally grammatical sentence but one which has a different meaning. Any other word order is unacceptable (although words such as *the* can be omitted in 'shorthand' forms, such as newspaper headlines). In Modern Greek, the equivalent sentence (transcribed into the Latin alphabet) is:

O skilos eida ti gata.

To say *the cat saw the dog* the sentence must become:

I gata eida to skilo.

When the dog is the 'subject' of the sentence it is *o skilos*; when it is the 'object' of the sentence it becomes *to skilo*. Similarly *i gata* changes to *ti gata* as it moves from subject to object. Thus in Greek if we take the first sentence and reverse the order of *o skilos* and *ti gata,* giving:

Ti gata eida o skilos.

the sentence still means that the dog saw the cat. The three elements of the sentence (subject, verb and object) can be placed in any order without altering the basic meaning. The morphology of the words (primarily but not exclusively their endings) makes clear who is doing what to whom. Word order merely emphasizes different elements. To get the effect of the last sentence above in English, we would need to completely re-cast it. Two possibilities are:

It was the cat that the dog saw.
As for the cat, the dog saw it.

The history of many Indo-European languages seems to be one in which a combination of complex morphological differences with fairly free word order has given way to smaller morphological differences accompanied by rigid word ordering. Classical Latin or Greek and Modern Russian stand at one end of this spectrum; Modern Greek somewhere in the middle; French and English at the other end.

Some definitions are needed. Words can often be divided into **morphemes**. Thus *inputs* (whether noun or verb) can be divided into three morphemes: *in* + *put* + *s*. Note that ALL THREE parts of the word are called morphemes: *put* is the **root** morpheme to which the **affixes** *in* and *s* are added. Affixes can be **prefixes**, **infixes** or **suffixes**, i.e. can be added to the start of a word, inserted into a word or added to the end of a word. In English, infixes are rare,[1] although they are common in other languages (Arabic for example).

Some authors define morpheme as the smallest unit of meaning. There are some objections to this definition, which is why I never use it.

- Units which alter meaning are not necessarily morphemes. The pluralization of *man* to *men* or *mouse* to *mice* (clearer in the phonemic representations /maʊs/ and /maɪs/)

[1] One example occurs in certain forms of swearing. Thus *You can stuff this uni-bloody-versity*. Note that the infix cannot occur anywhere in the word. *Univers-bloody-ity* is ungrammatical.

involves a phenomenon often called 'umlaut': a more-or-less regular change in the vowel phoneme in the root morpheme, common in Germanic languages. Such changes don't involve the addition or replacement of morphemes; we can't analyse *men* as being made up of two component morphemes. Nevertheless the change of vowel phoneme does change one aspect of meaning (singular to plural). The 'unit' involved here is smaller than a morpheme, so a morpheme isn't the smallest unit of meaning.

- Morphemes don't always alter meaning in any significant way. For example, in languages where adjectives change with gender, this may involve changing an inflectional morpheme (e.g. *nouv-eau* to *nouv-elle* in French). However, the adjective doesn't change meaning, since syntactic gender isn't consistently associated with a semantic feature. The newness of a new hat (*un chapeau nouveau*) isn't different from the newness of a new shirt (*une chemise nouvelle*). The inflectional morphemes here aren't units of meaning in any useful sense; they have a purely syntactic function.

To summarize: it is useful in some circumstances to view words as being composed from sequences of smaller parts. These are called morphemes. Morphemes may affect meaning, but equally may only show syntax changes.

The term 'word' is not sufficiently precise in the context of morphology. Are *inputs* and *input* the same word? What about *man* and *men*? One solution is to distinguish between a **word** and a **lexeme**. *Man*, *men*, *girl* and *girls* then constitute four different words but only two different lexemes. A lexeme is, roughly, what is entered into a dictionary.[2] *Men* is just the plural of *man*: no new meaning is implied other than 'plurality', so only one dictionary entry is required.

Corresponding to this distinction are the ideas of **inflectional** and **derivational** morphology. Consider the word *allow* which can be used in a sentence like:

We allow smoking.

Adding the inflectional morphemes *s*, *ing* and *ed* produces new words but the same lexeme:

He allows smoking.
We are allowing smoking.
Smoking is allowed.

In all of these sentences, the basic meaning of *allow* is preserved. On the other hand, adding the derivational morpheme *ance* produces a different lexeme, i.e. a different meaning:

**Smoking is an allowance.* (This does not mean that smoking is allowed.)

Adding inflectional affixes produces a new word but not a new lexeme: the basic meaning of the word is unchanged, as usually is its part of speech. Adding derivational affixes changes the word from one lexeme to another, changing the basic meaning and often also the part of speech.

In modern English, the few inflectional morphemes left are all suffixes, e.g.:

cat + s → cats
push + ed → pushed
push + ing → pushing

In writing, *-s*, *-'s*, *-s'*, *-ing* and *-ed* are the only regular inflectional suffixes (*-en* occurs in a small subset of verbs, e.g. *eat – eaten*, SAE *got – gotten*). The addition of *-ly* to an adjective to form an adverb is also regular, although it does change the part of speech, so is not strictly speaking inflectional.

Derivational affixes can be either prefixes or suffixes:

up + tight → uptight
tight + ness → tightness

More than one prefix and suffix can be used together, e.g.:

*un + dis + charge + ed → undischarged (*2 prefixes, 1 suffix*)*
*un + know + ing + ly → unknowingly (*1 prefix, 2 suffixes*)*

---

[2] Lexicographers often talk of 'head words' rather than 'lexemes'.

*anti + dis + establish + ment + ary + an + ism → antidisestablishmentarianism (*2 prefixes, 4 suffixes*)*[3]

It seems clear that NLP systems must be able to cope with inflectional morphology. In English, for example, we don't want to explicitly store the plural of every noun, since these are mostly very predictable. The relatively few exceptions can be stored separately, and rules used to generate the rest. The same applies to verbs, where the addition of *s*, *ing* and *ed* should be handled by morphological rules (although exceptions are somewhat more common).

Derivational morphology, on the other hand, is much less regular. Some verbs can be converted to nouns by adding *ance*, e.g.

allow + ance → allowance
utter + ance → utterance
deliver + ance → deliverance

However many cannot:

*express + ance → *expressance* (cf. *express + ion → expression*)
*derive + ance → *derivance* (cf. *derive + ate + ion → derivation*)
*develop + ance → *developance* (cf. *develop + ment → development*)

There are also phonological differences in the two kinds of affixing. For example, *sign* is pronounced [saɪn], following a rule of English that the sequence [gn] is not allowed within a syllable. Inflections of *sign* have the same pronunciation (e.g. *signed* is [saɪnd]) since the 'root' morpheme remains a full syllable. However, derived words may behave differently. Thus *signal* or *signature* split the root across syllables (e.g. *signal = sig-nal* → [sɪg nəl]).

Thus morphological rules seem likely to be more useful in processing inflectional affixes.

2 Written or Spoken?

In a language like English with its non-phonemic spelling system, different morphological rules are needed for the written and spoken languages. (French is not much better.) Consider the plurals of nouns. In the written language, the regular plural of nouns is formed by adding *s* or *es*. The regular rules are:

W1 Singular words which end in *s*, *z*, *sh*, *ch*, or *x* add *es*.
Examples: *bus/buses, whizz/whizzes, bush/bushes, church/churches, fox/foxes*.

W2 Singular words which end in consonant + *y* change the *y* to *i* and add *es*.
Examples: *cry/cries, fly/flies*.

W3 All other singular words add *s*.
Examples: *hop/hops, hope/hopes, play/plays, piano/pianos*.

(The last example illustrates one area of difficulty: words ending in *o* vary in whether they add *s* or *es*; e.g. *tomato* becomes *tomatoes*. A former Vice-President of the USA is notorious for wrongly 'correcting' a pupil's spelling of such a word!)

In the spoken language, the regular morphological rules are different:

S1 Words which end in alveolar or palatal fricatives or affricatives add [ɪz].
Examples: [bʌs]/[bʌsɪz], [dʒʌdʒ]/[dʒʌdʒɪz].

S2 Words which end in other unvoiced phones add [s].
Examples: [kæt]/[kæts], [kɒf]/[kɒfs].

S3 All other words add [z].
Examples: [dɒg]/[dɒgz], [dʌv]/[dʌvz], [rʌnə]/[rʌnəz].

In either case, words not covered by the rules can be marked in the lexicon as having irregular plurals (e.g. *man / men*). An alternative for some words is to mark them as being formed by different rules. Thus we can either mark *knife* as having the irregular written plural

[3] Historically, *establish* could be derived from *e + stable + ish,* which would make 3 prefixes and 5 suffixes!

knives or mark it as forming its plural by an 'irregular' rule:[4]

W4 Words ending in *ife* change the *f* to *v* and add *s*.
Examples: *wife/wives, knife/knives, life/lives*.

Note that what counts as 'irregular' depends entirely on the 'regular' rules used. Should we treat *piano* as regular (as above), which means that *tomato* is irregular, or should we treat *tomato* as regular and *piano* as irregular?

The boundary between morphological rules and phonological rules is 'fuzzy'. An alternative way of handling regular English plurals in the spoken language is to use one morphological rule followed by two phonological rules:

S1 To form the regular plural of an English word, add the phoneme /z/.

P1 /z/ → [ɪz] : alveolar or palatal fricative or affricative _ end of word

P2 /z/ → [s] : unvoiced phone _ end of word

Implementation of Morphological Rules

Implementing morphological rules can be achieved in various ways. Consider written English where only suffixes are involved. We can usefully write the morphological rules a little more formally than above. Those derived from W1 to W3 are:

s → *ses*	*z* → *zes*	*sh* → *shes*
ch → *ches*	*x* → *xes*	
C*y* → C*ies*	L → L*s*	

where L is any letter, C any consonant letter and the lhs and rhs of the rule are both word endings. The rules must be tried in order (why?).

A simple algorithm to add the plural *s* is given below. Note carefully that the rules are tried in order and only the first match used. (The algorithm is often easier to code if the rules are re-written to have a constant number of letters on the lhs.)

```
FUNCTION add_s(word)
  FOR each rule IN the set of rules LOOP
    IF the last letters of the word match the lhs of the rule THEN
      RETURN the first part of the word + the rhs of the rule
    END IF
  END LOOP
  RETURN word
END add_s
```

An obvious problem with this approach is that given the rules above and the word *man* it will return *mans*. To correct this we need a 'dictionary lookup':

```
FUNCTION plural(word)
  IF word is in the dictionary with an irregular plural THEN
    RETURN the irregular plural
  ELSE
    RETURN add_s(word)
  END IF
END plural
```

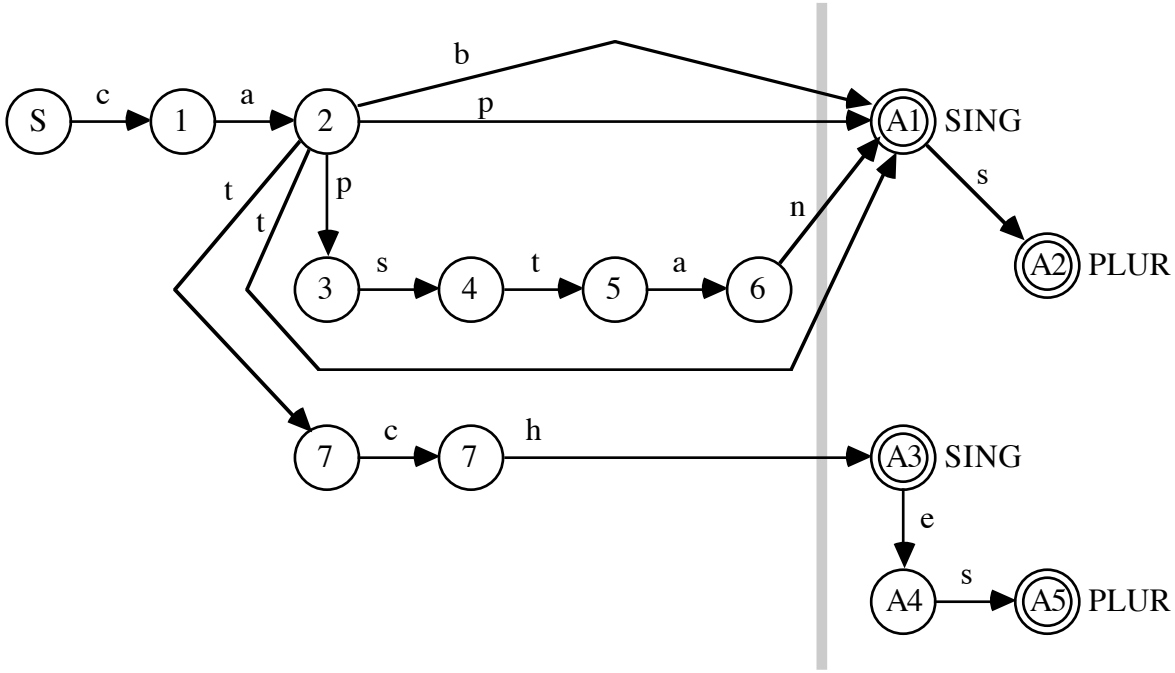
The reverse process – deriving the singular of a noun from its plural – is equally straightforward to code (see the Exercises).

Improving Efficiency

The general approach outlined above to handling the inflectional morphology of words can be implemented in many languages. In some cases, efficiency can be improved. For example, we could store words as a doubly-linked list of letters which would give direct access to the

[4] A different solution to this problem is discussed in the Answers to the Exercises.

end of the word in order to remove the suffix. A more efficient approach is to transform the entire lexicon plus morphological affixes into a transition network. The diagram which follows shows one way of storing the words *cab*, *cap*, *capstan*, *cat* and *catch*, together with the plural morphemes *s* and *es*.



We start from node S, and try to reach a terminal node (marked with a double circle), consuming the appropriate letter for each arc traversed. A successful traversal ends on a terminal node with no letters left. Thus the word *caps* leads from S (via 1, 2 and A1) to the terminal node A2. This node is marked PLUR showing that *caps* is a plural word. The network is 'non-deterministic' in that there may be a choice of arcs at a given node, so that backtracking is necessary if the chosen arc leads to an unsuccessful path. When carefully constructed, such networks can be highly efficient, both in space (since most arcs will be shared between many words) and in time (since following ONE path determines both whether the 'word' exists and its morphology – for example, any word which ends at A2 has the suffix *s* and is plural).

Writing the code for such a network by hand is tedious and error-prone, so they are usually 'compiled' from a more human-friendly form of lexicon. How this might be done is beyond the scope of this module.

Spelling dictionaries incorporated in word-processors are another application for networks of this kind. Here we merely wish to decide whether the word exists, rather analyse than its morphology, which can allow the use of more efficient DETERMINISTIC networks.

Exercises

1. The words in the first four sentences of this handout are listed below.

a	*example*	*of*	*such*
against	*feature*	*off*	*syntax*
and	*grammar*	*one*	*tend*
another	*have*	*ordering*	*the*
as	*includes*	*original*	*these*
been	*Indo-European*	*others*	*thought*
being	*is*	*Persian*	*to*
both	*language*	*retained*	*trade*
but	*languages*	*rigid*	*which*
complex	*less*	*rules*	*within*
components	*more*	*Russian*	*words*
elements	*morphology*	*sentences*	
English	*notable*	*some*	

   a) Find all the words with inflectional morphology and identify their morphemes.

   b) Choose FIVE words with derivational morphology and identify their morphemes.

2. Consider the following Modern Greek and English sentences or phrases:

O skilos eida ti gata.	*The dog saw the cat.*
I gata eida to skilo.	*The cat saw the dog.*
Oi skiloi eidan tis gates.	*The dogs saw the cats.*
Oi gates eidan tous skilous.	*The cats saw the dogs.*
To fayito tou skilou ...	*The dog's food ...*
To fayito ton skilon ...	*The dogs' food ...*
To fayito tis gatas ...	*The cat's food ...*
To fayito ton gaton ...	*The cats' food ...*

   The Greek *to fayito tou skilou* is literally *the food of-the of-dog*. Identify all the morphemes in the written forms of the Greek and English words for dog and cat, and draw up a table showing the correspondence between the Greek and English words.

3. In written English, the singular and plural possessives (genitives) are formed as follows:

   dog → *dog's / dogs'*
   man → *man's / men's*

   a) Write down in some appropriate format ALL the rules which govern the morphology of the possessive in written English.

   b) Write down the corresponding rules for spoken English.

4. Consider adding the morpheme *ed* to the end of written English verbs using the 'regular' rules. For example:

   jump + ed → jumped
   hope + ed → hoped
   cry + ed → cried
   play + ed → played
   hop + ed → hopped

   Note that in British English, the 'doubling rule' illustrated in the last example applies only to words ending C_1VC_2 (where C=consonant, V=vowel) and where C_2 is not *w*, *x* or *y*.[5] (Among other differences, in American English *l* is not doubled either.)

   knot + ed → knotted (ends in CVC)

---

[5] The 'doubling rule' is actually more complicated and has different contemporary usages. Stress is also important. The *r* is always doubled in *referred* but not in *levered*; *bussed* always spelt with double *s* but *focused* is increasingly not. Why?

refer + *ed* → *referred* (ends in CVC)
shout + *ed* → *shouted* (doesn't end in CVC)
show + *ed* → *showed* (C_2 = *w*)

Try to write all the rules. What algorithm is needed? What if it is required to find the past form of irregular verbs (e.g. *catch* → *caught* rather than **catched*)?

5. What algorithm should be used for the reverse process to that considered in Exercise 4, i.e. finding the present tense of a verb from its past? What will your algorithm do if its input is *catched*?

6. Write down a list of English NOUNS ending in each of the consonants given in the table in the Phones and Phonemes handout. (No English word ends in [h]. Only words derived from other languages, usually French, end in [ʒ] – e.g. *rouge*, *mirage*, *garage* – and in many English dialects, [dʒ] is normally used instead. Nouns ending in [ð] are rare; some examples are *lathe*, *scythe* and *tithe*.)

   a) In your speech, do the spoken plurals of the nouns in your list follow the rules S1 – S4 given above?

   b) Here are some pronunciations which I use for plurals (at least in fast, non-formal speech).

      cough [kɒf] – [kɒfs]
      loaf [loʊf] – [loʊvz]
      moth [mɒθ] – [mɒθs]
      mouth [maʊθ] – [maʊðz]

      Note that English spelling shows the 'irregular' plural of *loaf* (*loaves*) but can't do so for *mouth*. Are words like *loaf* and *mouth* irregular or is there a rule? (What is the plural of *bath* in dialects which pronounce it [bɑθ] and in dialects which pronounce it [bæθ]?)