

# Report on NLPA Examination 2006/07

Peter Coxhead

## Notes on some Examination Questions

I've only given notes where the answers are not straightforwardly from textbooks or handouts.

1. b) i) The change [ænd] → [ənd] → [ən] → [n] was given, so three rules are needed for the three steps. The first rule is that unstressed /æ/ becomes [ə]. So either of the two following steps could be chosen for the answer.

[ənd] → [ən] The most plausible rule is that a voiced stop after a nasal in the same position at the end of a word can be omitted. The formal rule would be something like

{stop, voiced, Posn} → ∅ : {nasal, Posn} \_

(Beyond the examination, but support for this rule is that the spelling *mb* at the end of English words always corresponds to the pronunciation [m], and that in most dialects of English the spelling *ing* at the end of a word is not [ɪŋg] but just [ɪŋ].)

[ən] → [n] The most plausible rule is that [ə] can be omitted when before a nasal. Formally

[ə] → ∅ : \_ {nasal}

(Beyond the examinations, but words like *bottom* or *button* are normally pronounced without a vowel between the [n] and the nasal.)

Any plausible answers were given credit, but they had to relate to the two remaining steps.

- ii) *handbag* is pronounced *hambag* because of the sequence /hændbæg/ → [hænbæg] (d omission rule from above) → [hæmbæg]. So the missing rule is that one that gets from [nb] → [mb]. This is nasal assimilation (covered in the handout):

{nasal} {stop, Posn, voiced} → {nasal, Posn} {stop, Posn, voiced} OR  
{nasal} → {nasal, Posn} : \_ {stop, Posn, voiced}

- c) The diagram was given to indicate things I didn't want you to have to remember OR to write about. Clearly there weren't marks for just writing out in words what the diagram says! I wanted to know that the 'features' *weren't* linguistic but things like (formant) frequencies, intensities, etc.; how the features were recognized as *linguistic* units (e.g. could use neural net or other pattern matching approaches); what the linguistic units were (phones or sub-phones e.g. split stop into closure + release); how language models could be used (via N-grams/Markov models, look at probability of sequences of phones and of words – can have specific language models for contexts, e.g. words found in business letters).
2. a) Quite a few candidates lost marks for not “indicating its relevance to NLP” for each of the three concepts.
- c) For (i), either use morphological analysis (*designing* is an inflectional variant and hence the same lexeme as *design*, whereas *designation* is a different lexeme), with the knowledge that in English the root morpheme has the same pronunciation in inflectional variants, or use look-up for all words.

For (ii), the problem is that the graphemes are identical. Only by doing syntax analysis can a TTS system realize that there are two lexemes, the noun and the verb, and then either use rules or lookup to decide on the pronunciation.

Either way, this illustrates the inter-dependence of levels of processing, which is the main consequence.

3. b) Best solution (?) is to add a variable to *eVerb* entries in lexicon to indicate valid complement type and modify the grammar accordingly. E.g.

ES → eNoun(N) EVP(N)

EVP(N) → eAux(N) eVerb(**CType**) EComp(**CType**)

EComp(**null**) → ∅

EComp(**sentence**) → *that* ES

eNoun(s) → *Max*

eNoun(p) → *people*

eAux(s) → *has*

eAux(p) → *have*

eVerb(**null**) → *eaten*

eVerb(**sentence**) → *learnt*

I marked your tree based on your grammar, whether it was right or wrong (at least if it was possible to draw a tree based on your revised grammar!).

- c) The broad idea is English sentence → English syntax tree → German syntax tree → German sentence.

Step 1 uses an English grammar + lexicon for morphology.

Step 2 needs to do TWO things: reorder the tree and 1:1 translation. The example shows that in the subordinate clause (sentence after *that*) the English syntax tree needs to be re-ordered (auxiliary and verb swapped) before or after 1:1 translation of the lexemes. Surprisingly few answers said this.

Step 3 puts the now German syntax tree back through a German grammar and lexicon. At this stage things like the correct gender, case, etc. in German will be fixed.

Problems are:

– impossible generally to translate lexemes 1:1 independently of the context (e.g. words like English *wood*, *bank* which are homographs – different lexemes written the same).

– required syntax structures in different languages are often not simple tree transforms of one another (e.g. *I like dancing* to *dancing pleases me* in Spanish, Greek).

– idioms, metaphors may simply not translate (e.g. *raining cats and dogs*).

4. a) I think this is fairly straightforward. One possibility is:  
*problem* [+ABSTRACT, -ANIMATE], *dog* [-ABSTRACT, +ANIMATE]  
*easy* [+ABSTRACT], *happy* [+ANIMATE], *eager* [+ANIMATE]  
Have a rule saying that features of noun & adjective must not clash, i.e. must not have same feature name with opposite sign. Look up features in lexicon and check this.
- b) This is more difficult, and was marked generously accordingly. If we keep features to the noun and adjective then based only on the examples given, in *The* noun *is* adj *to please*, the noun must be [-ABSTRACT, +ANIMATE] regardless of the adj. In *It is* adj *to please* noun, the same is true of the noun, but now the adj must select for +ABSTRACT.

An alternative might be to try to put features on the verb (here *please*).

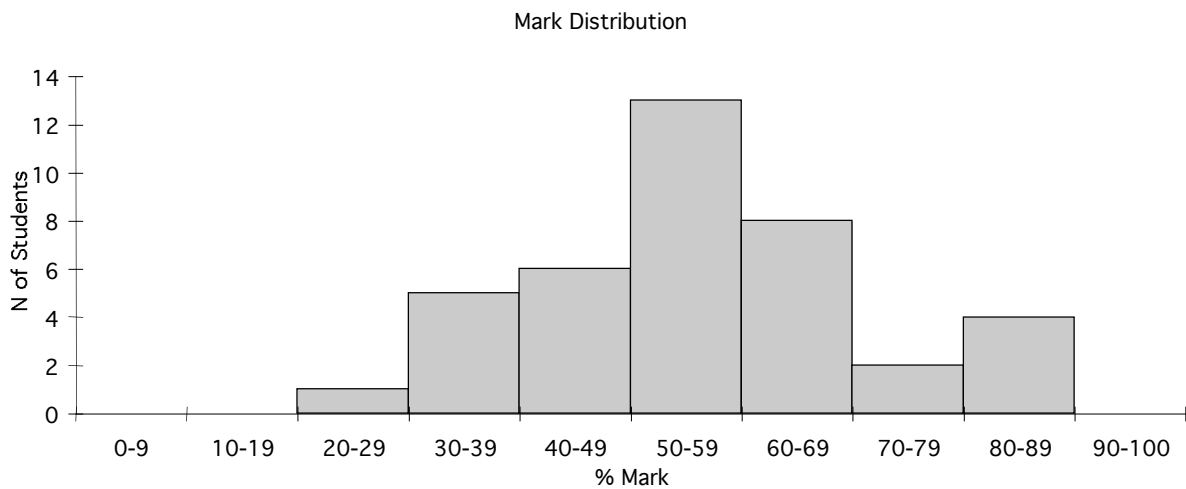
Some candidates said something like *The dog is eager to please* means that the dog is doing the pleasing whereas *The dog is easy to please* means that the dog is being pleased, and so different features are required. This is wrong unless a algorithmic way of deciding between the two interpretations can be given; it can't be used as the starting point for a computerized system otherwise.

## Examination Marks

Note: if you look at your marked script, the total is out of 60 as I marked each question out of 20.

Some students contacted me after the exam to say that it was a bit too long. I haven't found it easy to cut questions down to fit a 1.5 hour paper after many years of writing them for a 2 hour paper. When marking scripts, it did seem to me that some **better** candidates ran out of time on their last question (they often said this on the paper). There's no evidence that weaker candidates had this problem – they didn't know enough to run out of time! After some thought, I decided to scale that proportion of the exam mark which was above 50% up by 15%, i.e. for all  $\text{raw\_percentage} > 50$ ,  $\text{scaled\_percentage} = (\text{raw\_percentage} - 50) * 1.15 + 50$ . This results in a very modest increase in the mean mark on the exam from 54.8% to 56.0%. (The raw averages for the previous three years were 59%, 53% and 55%.)

The final percentage mark distribution on the examination only is shown in the histogram below. It's reasonably symmetrical around the average. 6 candidates had fail marks on the exam alone, 6 had first class marks.



## Module Marks

The total mark for the module is made up of 80% from the examination, 20% from the coursework. The distribution of the module marks is shown in the histogram below. 3 candidates failed; 6 had first class marks.

