Assignment 1

Given on: 22 January, 2019
Due date: 1 February, 2019

Take or make a text of 300 words. Annotate the text using the USAS semantic codes as shown in the example.

Semantic tagging

Beyond grammatical annotations, semantic annotation is an obvious next step. For example, semantic word-tagging can be designed with the limited (though ambitious enough) goal of distinguishing the lexicographic senses of same word: a procedure also known as `sense resolution'.

The ACASD semantic tagging system (Wilson and Rayson, 1993) accepts as input text which has been tagged for part of speech using the CLAWS POS tagging system. The tagged text is fed into the main semantic analysis program (SEMTAG), which assigns semantic tags representing the general sense field of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. *as a rule*), which are updated as new texts are analyzed. (Items not contained in the lexicon or idiom list are assigned a special tag, Z99, to assist in updating and manual postediting.) The tags for each entry in the lexicon and idiom list are arranged in general rank frequency order for the language. The text is manually pre-scanned to determine which semantic domains are dominant; the codes for these major domains are entered into a file called the `disam' file and are promoted to maximum frequency in the tag lists for each word where present. This combination of general frequency data and promotion by domain, together with heuristics for identifying auxiliary verbs, considerably reduces mistagging of ambiguous words. (Further work will attempt to develop more sophisticated probabilistic methods for disambiguation.) After automatic tag assignment has been carried out, manual postediting takes place, if desired, to ensure that each word and idiom carries the correct semantic classification (SEMEDIT). A program (MATRIX) then marks key lexical relations (e.g. negation, modifier + adjective, and adjective + noun combinations). The following is an example of semantic word-tagging, taken from the automatic content analysis project at Lancaster:

EXAMPLE OF SEMANTIC TAGGING

```
PPIS1     I           Z8
VV0       like        E2+
AT1       a           Z5
JJ        particular  A4.2+
NN1       shade       O4.3
IO        of          Z5
NN1       lipstick    B4
```

In this fragment, the text is read downwards, with the grammatical tags on the left, and the semantic tags on the right. The semantic tags are composed of:

- an upper case letter indicating general discourse field
- a digit indicating a first subdivision of the field
- (optionally) a decimal point followed by a further digit to indicate a finer subdivision
- (optionally) one or more `pluses' or `minuses' to indicate a positive or negative position on a semantic scale

For example, A4.2+ indicates a word in the category `general and abstract words' (A), the subcategory `classification' (A4), the sub-subcategory `particular and general' (A4.2), and `particular' as opposed to `general' (A4.2+). Likewise, E2+ belongs to the category `emotional states, actions, events and processes' (E), subcategory `liking and disliking' (E2), and refers to `liking' rather than `disliking' (E2+).

The semantic annotation is designed to apply to open-class or `content' words. Words belonging to closed classes, as well as proper nouns, are marked by a tag with an initial Z, and set aside from the statistical analysis.

For more information, see USAS web page, Thomas and Wilson (1996) and Garside and Rayson (1997).

Top level semantic codes:

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			