Using the Historical Thesaurus Semantic Tagger

The Historical Thesaurus Semantic Tagger (HTST) can be used to label lexical items in running text with codes based on the semantic categories of the *Historical Thesaurus of English*. The Java GUI allows moderate sized texts (up to the average modern book length of around 100,000 words) to be tagged. For texts larger than this, it is advisable to consult Dr Paul Rayson at UCREL (Lancaster University) on the possibility of using their servers for the job.

The HTST's interface takes the form of two parallel windows. Text can be loaded into the left window, tagged by the programme, and the results of the tagging process returned in the right window. The **File** menu gives options for loading text into the windows, saving text from the windows, and clearing them. It is possible to either type or copy-and-paste plain text into the windows for tagging, and there is a drop-down menu for the type of character set used should this need to be changed.

Tagging itself is initiated through the **Tools > English Semantic Tagger > Tag text in left window** menu option. There is also the option here to select a file for tagging rather than loading it into one of the windows. Important adjustments to the tagging process are found in the **Setting** menu. These allow the user to specify a **date range** for the text, helping the tagger to filter out senses which are likely to be anachronistic for the text in question. The spelling normaliser VARD can be used as part of the tagging process, and this can be switched on or off from the **Use VARD** option in the Setting menu. When active, VARD replaces variant spellings of words in the text with a normalised spelling in the ouput, which is then used as the basis on which the HTST assigns a semantic code to that word.

Reading the HTST Output

The output of the HTST is arranged in rows with one row per word of the input text. If manual analysis is preferred, it is often easiest to read the output if it has been saved as a .txt or .csv file which can then be opened in a spreadsheet package. The columns produced in the output are:

- #TOTEN – An individual token, usually a word or a punctuation mark.
    - *N.B. If VARD is active, the original token will be replaced in this column by the normalised spelling which results from the VARD analysis.*

- LEMMA – The lemmatised (or 'dictionary') form of word tokens.

- POS – The part of speech of the token (e.g. noun, verb, adjective). A list of the possible part of speech tags can be found at http://ucrel.lancs.ac.uk/claws8tags.pdf)

- SEMTAG1 – A semantic tag for the token based on the thesaurus used by the USAS tagger. This was also developed by UCREL, and the list of thesaurus categories can be found at http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf

- MWE – a numerical formula which indicates the structure of a 'multi-word expression' where this is appropriate.

- SEMTAG2 – a semantic tag placing the word sense in the hierarchy of the *Historical Thesaurus of English*. These can be found through the *Thesaurus'* website: www.glasgow.ac.uk/thesaurus

- SEMTAG3 – a semantic tag placing the word sense within the 'thematic' category set, also based on the hierarchy of the *Historical Thesaurus of English*, but employing a 'human scale' structure which allows important concepts to be identified more readily. The listing of thematic categories along with a readme file explaining their creation and structure can be found on the SAMUELS website: http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/