

L113: Course Work I

Simone Teufel

7 October 2011

Consider the corpus examples for “interest” and “line” (files `interest.html` and `line.html` on course webpage). They each show 150 sentences of the target words, extracted at random from the British National Corpus.

Your task is to decide how many senses the target words have, using the corpus evidence as your input. You are allowed to choose which target word you prefer.

- Start by clustering the sentences where you think senses are identical. Give each cluster a temporary title.
- Cluster as many sentences as you think are necessary for determining how many senses there are. You may not need to cluster all sentences, but you can if you want to. After clustering your sentences, make sure that different clusters really carry different senses (no merging possible), and that each cluster really only represents one sense (no splitting possible). In cases where you are not certain, apply the tests from the lecture.
- Write a definition for each sense, and select one representative sentence from the list for each sense (Sentence IDs are enough).

Submit a report that includes your list of senses with definitions. Indicate how you arrived at the list, including how many sentences you clustered. For each pair of similar senses, justify that they are really different. Apply at least one ambiguity test at least once, in order to demonstrate your use of the tests. Maximum length of report: 4 pages A4.

Submission deadline: Friday 14 October, beginning of lecture.