Adding linguistic information to parsed corpora

Susan Pintzuk
University of York

No matter how comprehensively corpus builders design their annotation schemes, users frequently find that information is missing that they need for their research, and so they must add it on their own. In this methodological talk I discuss and illustrate five methods of adding linguistic information of all types (lexical, phonological, morphological, syntactic, semantic, discourse) to corpora that have been morphosyntactically annotated (=parsed) in the style of Penn treebanks, and the advantages and disadvantages of each method. These five methods are the following:
1) adding information to the ur-text; 2) inserting 'CODE' nodes into the token structure; 3) embedding information in coding strings; 4) modifying node labels and structure; and finally, 5) moving token information from the corpus into spreadsheets. Methods 1 and 2 are necessarily manual, while methods 3, 4, and 5 involve a combination of manual and CorpusSearch functions and tools. Of course the main goal, regardless of method, is to record additional information within the corpus data so that the information can be retained through further searches and data processing.

Methods 1 and 2 are the simplest and, being manual, the most prone to error. They consist of adding information to the two areas of CorpusSearch output that are reproduced each time CorpusSearch is run: the token ur-text, which contains the text and token ID without any annotation, and the token structure, which may dominate CODE nodes containing text markup. The main difference between the two methods is that material internal to the ur-text is not visible to CorpusSearch and therefore not searchable, while CODE nodes are part of the token structure and therefore can be used within CorpusSearch queries.

Method 3, the construction of coding strings, is the traditional and perhaps most widely used method of adding information to corpus data. Coding strings had their origin in quantitative sociolinguistic research and were used decades before the creation of parsed corpora. Coding strings are strings of characters, each representing a linguistic or extralinguistic variable, which are inserted in the tokens of a corpus file.

The CODING feature of CorpusSearch can be used to construct coding strings, which may be manually extended to encode information that is not represented by the morphosyntactic annotation. Coding strings are part of the token structure and therefore may be searched and manipulated by CorpusSearch; coding strings may also be used as input to software for statistical analysis.

Method 4, the modification of corpus annotation, may be done manually, but it is much more efficient (and safe) to use the corpus-revision tool of CorpusSearch. This tool enables the addition, deletion, or modification of annotation in the corpus, including not only node labels but also structure.

Any search that can be made using CorpusSearch can act as the basis for corpus revision; the output of corpus revision is a new version of the corpus.

Finally, Method 5 copies coding strings from a corpus into a spreadsheet, the content of which may be ordered, manipulated, and displayed in ways that corpus data cannot be. For example, the data in the cells of a spreadsheet can be interpreted as numbers and used for simple calculations like totals, means, and frequencies; in contrast, the content of coding strings within a corpus are characters, not numerical values, and cannot be used in this way. As another example, the token itself and the preceding and following contexts can be added to a spreadsheet, while this is not possible within the corpus. Method 5 provides perhaps the most flexible way of working with and analyzing corpus data, but it should be used with caution, for at least two obvious reasons: it involves manual manipulation of the data, and therefore is prone to error; in addition, it is not always possible to go backwards from a spreadsheet to a corpus format.