

## B SOURCE, PREDICATE, OPINION extraction algorithm

# Sociological Tokenizer

1. Find complement clause(s) in the dependency parse of a sentence, i.e., sub-tree(s) whose root has the dependency label “ccomp” (= OPINION);
2. Get head(s) of the complement clause(s), which correspond to the main verb that syntactically embeds the comp. clause (= PREDICATE); get children of the PREDICATE with the dep. label “prt” (particle) in cases of multi-token verbs, e.g. *point out*;
3. To find the SOURCE, first check if the PREDICATE token is a participle (e.g., “a researcher, **warning** that [...]”—if yes, then find the head noun, otherwise, look within all children of PREDICATE and find the syntactic subject (token with the “nsubj\*” dependency label). In some cases, the head noun/syntactic subject may have the dependency label “relcl”, indicating that it’s inside a relative clause (e.g., “[...], **who** warns that”)—in this case, the true SOURCE is the antecedent of the relative pronoun, which we fetch by getting the head of the relative pronoun;
4. Get additional modifiers of SOURCE, PREDICATE and OPINION by recursively retrieving their children.

