REVISED ANNOTATION GUIDELINES

Paraphrase Status

- Sentences that are judged to be paraphrases should have at least one content word aligned (either sure or possible); if it's not possible to align any content words, then the sentences should not be considered paraphrases.

Multiple Occurrences of a Word or Phrase

- When a word or phrase appears twice in a question, align the one in most direct correspondence in terms of syntax and word order and mark the other one as possible.

Time

- With time reference, use possible alignments when different event times are implied, not just when different tenses are used: for example, present progressive and present perfect should be annotated with possible alignments.

Direct Substitution

- Use sure alignments when phrases are directly substitutable (in both directions) despite differences in syntax (e.g. for and to help with).

- Use sure alignments with hypernyms if they are fully substitutable in context, i.e. no important meaning is lost (e.g. hi for good afternoon).

Function Words

- Don't align function words (e.g. articles, WH-question words, and NPIs) that are modifying unaligned content words. Possessive pronouns should be aligned if they have the same reference.

- Use possible alignments when a pronoun is aligned to a full definite noun phrase (e.g. it for the pain).

- Prepositions that are only serving a syntactic function but don't affect the meaning should be left unaligned.

Verb Clusters

- When all the auxiliaries in a verb cluster are the same in the source and target, they should be singly aligned rather than block aligned (consistent with the Edinburgh corpus, though this detail is not spelled out in their guidelines).

- When a verb cluster in the source and target are of the same length and the main verb is inflected in the same way, they should be singly aligned rather than block aligned (e.g. singly align will be priced with would be priced, as in the Edinburgh corpus, though again this detail is not spelled out in their guidelines, and it's unclear how consistently such cases are handled).

- The verb have with a condition, such as have nausea, should be treated as a light/support verb construction and be block aligned with its equivalent in paraphrases (e.g. with be nauseous). Conversely, in non-paraphrases, have should not be aligned if the have + condition construction does not have an equivalent, even when there is an identical form of have (e.g. don't align have nausea with have STDs at all).

- Align verb clusters with adjectives or adverbs that provide the same time reference using possible alignments (e.g. have had for previous) when there is no corresponding modifier for the adjective or adverb (e.g. just align in the past with previous if such a modifier is present).

- Number agreement should be treated as a minor syntactic divergence and therefore possible alignments should be used (e.g. with the verbs make and makes).

- Don't align words that describe different things, including verbs that describe different events (even if they're the same vague verb like doing). A specific property or location of nouns (e.g., pain in legs vs. pain in back) is not necessarily enough to make them different in this way.

- Prepositions that do affect meaning like phrasal verbs should be aligned.

Tokenization and Spelling Errors

- If there's a typo that affects the tokenization (e.g. a run-on error), then it needs to be fixed before the annotation can be done. The tokenization error should be noted in the comments. Otherwise, if the intended word should be aligned, align the typo as a possible alignment.