A.2 The task

Paraphrasing stands for sameness of meaning between different wordings. For example, the pair of sentences in (a) are different in form but have the same meaning. Our **paraphrase typology (ETPC)** classifies paraphrases according to the linguistic nature of this difference in wording.

a) John said "I like candies"/John said that he liked sweets.

b) John said "I like candies"/John said that he liked onion.

The task described in these guidelines consists of annotating a Paraphrase Identification corpus (MRPC) with the Extended Paraphrase Typology (EPT). A Paraphrase Identification corpus contains textual paraphrase pairs (ex.: (a)), as well as textual non-paraphrase pairs (ex.: (b)). Our annotation task consists of two sub-tasks:

Annotating atomic paraphrases within textual paraphrase pairs (a) and textual non-paraphrase pairs (b). The textual pairs are generally complex in the sense that they contain multiple atomic paraphrases. We call these atomic paraphrases paraphrase phenomena and they are what should be annotated with the typology. The paraphrase pair in (a) contains two paraphrase phenomena: the direct/indirect style alternation and a synonymy substitution.

Annotating atomic non-paraphrases within textual non-paraphrase pairs (b). The non-paraphrase pair in (b) contains one atomic non-paraphrase: the substitution of "candies" with "onion".

In the annotation process, three main decisions should be made:

1) determine whether a candidate pair is a textual paraphrase (Section A.2.1)

2) If **non-paraphrase**, determine the key differences between the two texts:
   - choose the tag that best describes the phenomenon behind each difference (Section A.2.2)
   - determine the scope of every atomic non-paraphrase (Section A.2.3)

3) Determine the similarities between the two texts:
   - choose the tag that best describes the phenomenon behind each similarity (Section A.2.2)
   - determine the scope of every atomic paraphrase (Section A.2.3)

A.2.1 Is This a Paraphrase Pair

The first step in the annotation process is determining whether a candidate paraphrase pair is actually a paraphrase. We consider paraphrases those pairs having

the same or an equivalent propositional content. We consider non-paraphrases those pairs that have substantial difference in the propositional content. For example, a) will be annotated as “paraphrases”, while b) will be annotated as “non-paraphrases”.

a) Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

b) Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for $2.5 billion.
Yucaipa bought Dominick’s in 1995 for $693 million and sold it to Safeway for $1.8 billion in 1998.

Since the Extended Paraphrase Typology (ETP) can annotate atomic paraphrases (similarities) as well as atomic non-paraphrases (dissimilarities), both textual paraphrases and textual non-paraphrases will be subsequently annotated with the paraphrase typology. The subsequent annotation with paraphrase types will allow for distinguishing between paraphrase and non-paraphrase fragments within these sentences.

A.2.2 The Tagset

Our tagset is based on the Extended Paraphrase Typology shown in Table A.1. It is organized in seven meta categories: “Morphology”, “Lexicon”, “Lexico-syntax”, “Syntax”, “Discourse”, “Other”, and “Extremes”. Sense Preserving (Sens Pres.) shows whether a certain type can give raise to textual paraphrases (+), to textual non-paraphrases (-), or to both (+ / -). The typology contains 25 atomic paraphrase types (+) and 13 atomic non-paraphrase types (-).

The subclasses (morphology, lexicon, syntax and discourse based changes) follow the classical organisation in formal linguistic levels from morphology to discourse. Our paraphrase types are grouped in classes according to the nature of the underlying linguistic mechanism: (i) those types where the paraphrase arises at the morpho-lexicon level, (ii) those that are the result of a different structural organization and (iii) those types arising at the semantics level. Although the class stands for the trigger change, paraphrase phenomena in each class can entail changes in other parts of the sentence. For instance, a morpho-lexicon based change (derivational) like the one in (a), where the verb *failed* is exchanged for its nominal form *failure*, has obvious syntactic implications; however, the paraphrase is triggered by the morphological change. A structure based change (diathesis)

like the one in (b) entails an inflectional change in *hear/was heard* among others. Finally, paraphrases in semantics are based on a different distribution of semantic content across the lexical units with, on many occasions, a complete change in the form (c).

Table A.1 Extended Paraphrase Typology

ID	Type	Sense Pres.
	Morphology-based changes	
1	Inflectional changes	+ / -
2	Modal verb changes	+
3	Derivational changes	+
	Lexicon-based changes	
4	Spelling changes	+
5	Same polarity substitution (habitual)	+
6	Same polarity substitution (contextual)	+ / -
7	Same polarity sub. (named entity)	+ / -
8	Change of format	+
	Lexico-syntactic based changes	
9	Opposite polarity sub. (habitual)	+ / -
10	Opposite polarity sub. (contextual)	+ / -
11	Synthetic/analytic substitution	+
12	Converse substitution	+ / -
	Syntax-based changes	
13	Diathesis alternation	+ / -
14	Negation switching	+ / -
15	Ellipsis	+
16	Coordination changes	+
17	Subordination and nesting changes	+
	Discourse-based changes	
18	Punctuation changes	+
19	Direct/indirect style alternations	+ / -
20	Sentence modality changes	+
21	Syntax/discourse structure changes	+
	Other changes	
22	Addition/Deletion	+ / -
23	Change of order	+
24	Semantic (General Inferences)	+ / -
	Extremes	
25	Identity	+
26	Non-Paraphrase	-
27	Entailment	-