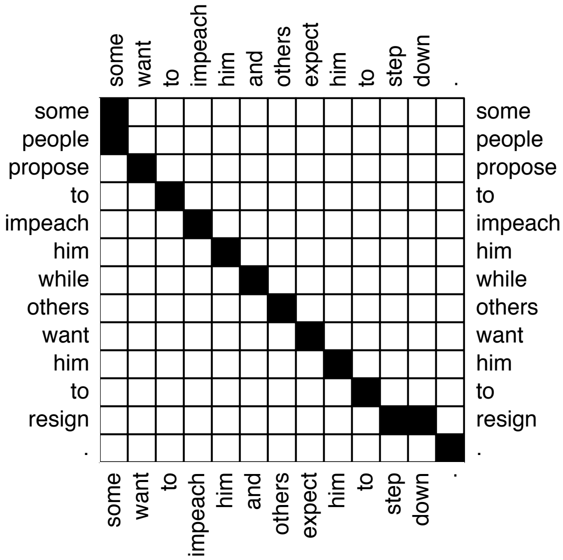
Annotation Guidelines for Paraphrase Alignment

Chris Callison-Burch Trevor Cohn Mirella Lapata

December 4, 2006

1 Introduction

You will be given pairs of sentences which are *paraphrases* of each other because they convey the same meaning but are worded differently. Your task will be to show which parts of the sentences are in correspondence by aligning them on a word-by-word basis. Here is an example of an alignment that we would like you to produce:



The dark squares in the grid indicate which words and phrases are in correspondence between the two sentences. These correspondences include simple cases where words are identical, but also more interesting cases where one word is replaced by another. The elements that correspond between the two sentences are the following, with re-wordings in bold: (***some***↔***some people***), (***want***↔***propose***), (*to*↔*to*), (*impeach*↔*impeach*), (*him*↔*him*), (***and***↔***while***), (*others*↔*others*), (***expect***↔***want***), (*him*↔*him*), (*to*↔*to*), (***step down***↔***resign***), (.↔.). Note that some correspondences are approximate (e.g., *want*↔*propose*) in that they convey slightly different semantics.

Your task will be to create alignments similar to the one above. To do this, we have provided tool on the web at `http://demo.linearb.co.uk/paraphrases/` which displays pairs of sentences in grids, and allows you to click on squares in the grid to indicate correspondences. Clicking on a square again changes the colour to grey, which indicates a *possible* alignment (more on this later). Clicking on a word on either axis flags the corresponding row or column as *new information* (more on this later). To save you some amount of effort the sentences have already been aligned automatically, but these automatic alignments are frequently inaccurate, so you will need to correct them carefully.

The basic instructions that we would like you to adhere when aligning sentences to are listed here, with more detailed instructions given in Sections 2 and 3:

1. The web tool allows you to mark two types of correspondences: *certain alignments*, which are indicated with black squares, and *possible alignments* which are indicated with gray squares. In general you should try to produce certain alignments. Section 3 gives a number of examples of where possible alignments might be needed.

2. In general you should prefer smaller alignments whenever possible. However, you are allowed to create *one-to-many alignments* where a single word in one sentence corresponds more than one word in the other, as

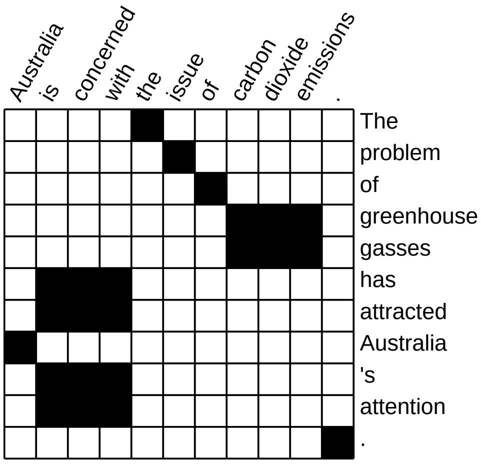
with (*step down*↔*resign*) or (*some*↔*some people*). You are also allowed to create *many-to-many alignments* where a sequence of words in one sentence corresponds to a sequence of words in the other. One-to-many and many-to-many alignments *do not* have to be marked as possible.

3. You should try to align all of the words in the sentences. Sometimes there will be information in one sentence which is not present in the other. In this case you should this information unaligned.

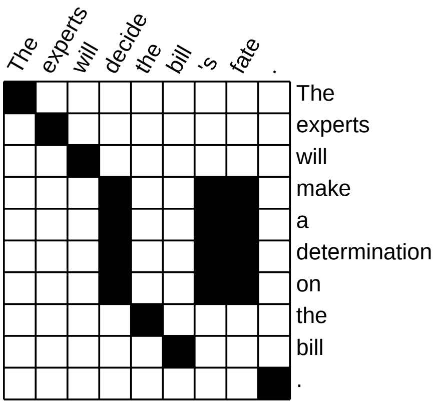
2 Aligning sentence pairs

We do not have a strict definition of what constitutes a correspondence. As a rule of thumb align words or phrases $a \leftrightarrow b$ in two sentences (A,B) whenever the words a can be substituted for b in B, or vice-versa. This relationship should hold within the context of the sentence pair in question: the relation $a \leftrightarrow b$ need not hold in general contexts. Trivially this definition allows for identical word pairs.

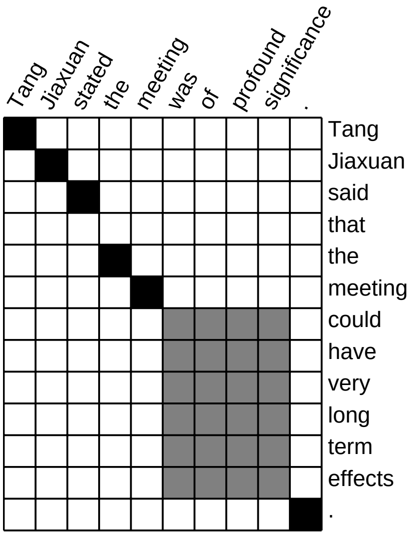
The choice of aligning words or blocks can be subjective. In general, you should align the smallest possible sensible unit, which will be individual words in the majority of sentences. When the same concept has been realised in the two sentences with quite a different construction, and aligning the two word-for-word becomes infeasible, then a many-to-many block should be used. For example, a block alignment is used in the following sentence pair:



The phrase *carbon dioxide emissions* corresponds to *greenhouse gasses* but cannot be aligned in a one-to-one fashion. Note that we do not use a many-to-many block for *Australia is concerned with* and *has attracted Australia 's attention* since it is possible to decompose the two phrases into smaller alignments. Specifically, we align *Australia* ↔*Australia* and *has attracted* <gap> *'s attention* ↔*is concerned with.* In general, discontinuous alignments should be preferred over many-to-many alignments. Another example of a discontinuous alignment is given below where *make a determination on* is aligned with *decide* <gap> *'s fate.*



Sometimes there are phrases across sentences which have only very loose correspondence. In these cases we have a special marker for "possible" alignments. These are shown as grey squares, and can be created by clicking twice on a white square. For example in the following sentence pair:

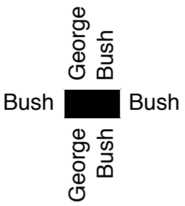


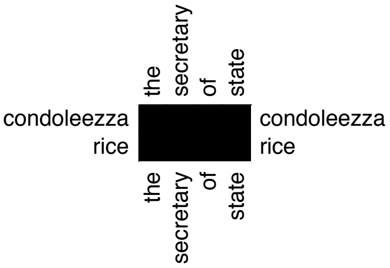
You might think that *could have very long term effects* was only loosely in correspondence with *was of profound significance*, in which case you should use the gray squares to mark a possible alignment. Possible alignments should also be used to mark significant changes in syntax but the words denote a similar concept. For example in cases where two words have the same stem but are expressed with different parts of speech, e.g., *co-operative* and *cooperation*; or when two verbs are used that are not synonyms, e.g., *is/marks* in *this is also* and *this also marks*; and the changing of determiners is observed.

3 Special Cases

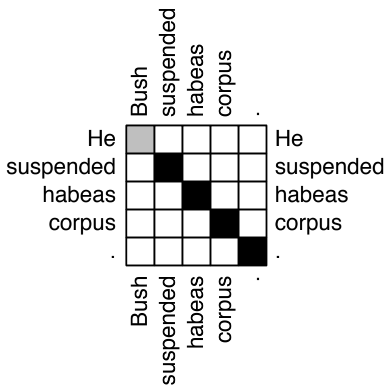
This section describes some general heuristics to decide how to make alignments in cases of ambiguity. The rules described here are fairly general and are simple to remember and rely on.

Named Entities Named entities denote people, places, organisations, etc. They are often expressed by proper noun NPs (e.g., *George Bush*, *the United Nations*) but can also be attested as definite descriptions (e.g., *the prime minister*, *the president of the organisation*). When two sentences have identical named entities, then these should be aligned on a word-by-word basis. For example for the pair *George Bush won* and *George Bush was victorious* you should produce the following alignments: *George* ↔ *George*, *Bush* ↔ *Bush* and *won* ↔ *was victorious*. When a sentence pair has two named entites that refer to the same individual, organisation, or location but differ in surface form, then you should use block alignments. So, *Bush* should be block aligned with *George Bush* and the *the secretary of state* with *Condoleezza Rice* in the example below. You should use block alignments even if the two entities differ with respect to one word (e.g., *George Bush* vs. *George W. Bush*). Finally, named entities can be recursive consisting of several named entities. In such cases, we recommend that you break the named entity into sub-parts each corresponding to one entity, instead of using a block alignment for the whole phrase. For example, if you were given the entites *Condoleezza Rice of the United States* and *Ms Rice*, you should block align *Condoleezza Rice* with *Ms Rice*, and leave *of the United States* unaligned.

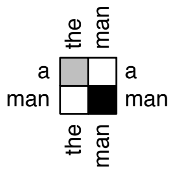


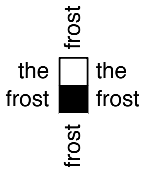


Pronominal reference When a pronoun is introduced in place of a name or a title, mark it is a possible alignment.

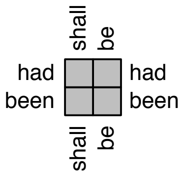


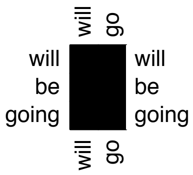
Determiners and Prepositions In cases where two NPs (PPs) have the same head but different determiners (prepositions) (e.g., *a man* vs. *the man*, *a ceremony* vs. *their ceremony*), the determiners should be aligned as possible. When an NP is missing a determiner altogether, then the determiner should be left unaligned (e.g., *frost* vs. *the frost*).





Tenses Use possible block alignments in cases where the same verb is attested with different tenses (e.g., *shall be* vs. *had been*). Use certain block alignments when the tense is the same but the aspect differs.

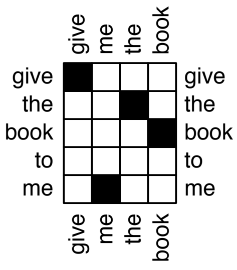




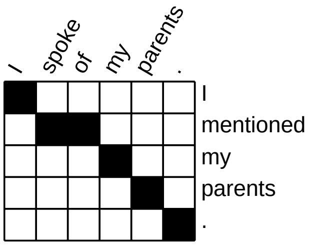
***To*-Clauses** When a *to*-clause (e.g., *to applaud*) corresponds to a verb in future tense (e.g., *will praise*), *to* should be aligned with the modal denoting the future tense (e.g., *will*). When the verb does not have a future tense, *to* should be left unaligned (e.g., *to have read* vs. *has read*).

Verb Complexes Verbs accompanied by modals and auxiliaries (e.g., *will sleep*, *might have done*, *could be eating*) should be considered one syntactic and semantic unit. Use certain many-to-one alignments in cases where such verb complexes in one sentence correspond to a single word in the other sentence (e.g., *was developed* ↔ *developed*, *will sleep* ↔ *sleep*).

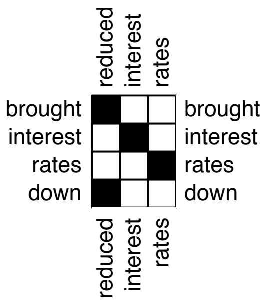
Subcat frames In cases where there the same verb is attested with differing subcat frames (e.g., *give to me* vs. *give me*), you should leave any extra material introduced by the different subcat frame unaligned.



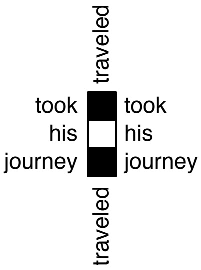
Use a many-to-one alignment when a verb with a prepositional complement paraphrases a verb that takes a NP complement.



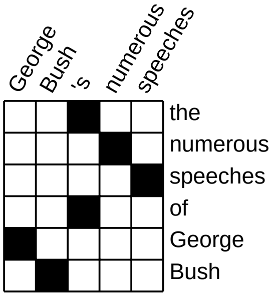
Phrasal Verbs Use a one-to-many alignment, in cases where a phrasal verb paraphrases a non-phrasal verb (e.g., *take up your father* vs. *bring your father*). In general phrasal verbs often take their meaning from the accompanying particle and should be considered as belonging to the verb. This can be a discontinuous alignment.



Support Verbs Use a one-to-many alignment in cases where a verb corresponds to a support verb construction (e.g., *took his journey* vs *traveled*, *had developed* vs *developed*). In other words, consider the support verb construction as one semantic unit.



Genitives You should align the *'s* maker with the word *of* and any determiner that is also introduced. This is shown below:



Typos and Approximate Correspondences The automatic alignment will probably go wrong when there are typos. For example, *help meet* instead of *helpmate*. This is also true when there are different surface forms for the same concept, e.g., *asshur* and *assyria*. The latter case is straightforward, the two terms should be aligned together. In the former case, the alignment should be indicated with possibles. If the sentence is particularly bad, you should communicated this to us, and we'll fix it.

Also there may be cases where phrases contain numbers that are not identical, e.g., *0.7 litres* vs. *1.5 litres* or *fifty-five percent* vs. *fifty percent*. Such alignments should be indicated with possibles. Where the units differ, such that the amounts are roughly equivalent, the phrase should be block-aligned.

Repetition In cases where the same entity is represented in a sentence multiple times, introduce multiple alignments. In the sentence pair below, you should mark one of the items as a “possible” alignment. The same holds not only for proper names (e.g., *the UN*) but also for pronouns, verbs, and common nouns.

The	UN	condemned	North	Korea	and	imposed	sanctions	.	
■									The
	■								UN
		■							condemned
			■						North
				■					Korea
					■				and
▨									the
	▨								UN
						■			imposed
							■		sanctions
								■	.

4 Rules of Thumb

While annotating keep in mind the following rules of thumb:

1. Try to annotate as much as possible.
2. Use block alignments when you are uncertain about which alignment to use.
3. Block align only named entities differing in surface form.
4. Mismatching pronouns should be marked as possibles.
5. Repeated instances of words or phrases should be marked as possible.

5 The Annotation Tool

The annotation will proceed on a sentence-by-sentence basis. Remember that you will be given automatically obtained alignments, so the sentences will be partially annotated. Once you review the sentence pairs and make whatever changes you see fit, save your alignments by pressing the *Submit* button. Then the next sentence pair will appear. You can also browse the data using the *<–Alignment–>* button at the beginning of the browser. You can move backwards by clicking on <– and forward by clicking on –>. It is important to remember that changes to already existing alignments can only be saved if you click on the *Submit* button. Finally, you can go to a sentence pair of your choice by pressing the *Linear B* button. This will allow you to type in the particular sentence pair number you wish to see.