What is a paraphrase?

Paraphrases are sentences or phrases that convey the same meaning using different wording. While the logical definition of paraphrases requires strict semantic equivalence, linguistics accepts a broader, approximate, equivalence, thereby allowing far more examples of “quasi-paraphrase”. But approximate equivalence is hard to define. Thus, the phenomenon of paraphrases, as understood in linguistics, is difficult to characterize. In this paper, we list a set of 25 operations that generate quasi-paraphrases. We then empirically validate the scope and accuracy of this list by manually analyzing random samples of two publicly available paraphrase corpora.

We provide the distribution of naturally-occurring quasi-paraphrases in English text.

1. Introduction

Sentences or phrases that convey the same meaning using different wording are called paraphrases. For example, the sentences (1) and (2).

1. The school said that their buses seat 40 students each.
2. The school said that their buses accommodate 40 students each.

Paraphrases are of interest for many current NLP tasks, including textual entailment, machine reading, question answering, information extraction, and machine translation. Whenever the text contains multiple ways of saying “the same thing”, but the application requires the same treatment of those various alternatives, an automated paraphrase recognition mechanism would be useful.

One reason why paraphrase recognition systems have been difficult to build is because paraphrases are hard to define. While the strict interpretation of the term paraphrases is quite narrow since it requires exactly identical meaning, in linguistic literature, paraphrases are most often characterized by an approximate equivalence of meaning across sentences or phrases. De Beaugrande and Dressler (1981) define paraphrases as “Approximate conceptual equivalence among outwardly different material”. Hirst (2003) defines paraphrases as “Talk(ing) about the same situation in a different way”. He argues that paraphrases aren’t fully synonymous: There are pragmatic differences in paraphrases, i.e., difference of evaluation, connotation, viewpoint etc. According to Mel’cuk (2012) “For two sentences to be considered paraphrases, they need not be fully synonymous: it is sufficient for them to be quasi-synonymous, that is, to be mutually replaceable salva significatione at least in some contexts”. He further adds that approximate paraphrases include implications (not in the logical sense, but in the everyday sense). Taking an extreme view, Clark (1992) rejects the idea of absolute synonymy by saying “Every two forms (in language) contrast in meaning”. Overall, there is a large body of
In this paper, we take a broad view of paraphrases. To avoid the conflict between the notion of strict paraphrases as understood in logic and the broad notion in linguistics, we use the term *quasi-paraphrases* to refer to the paraphrases that we deal with in this paper. In the context of this paper, the term “paraphrases” (even without the prefix “quasi”) means “quasi-paraphrases”. We define quasi-paraphrases as “Sentences or phrases that convey approximately the same meaning using different words”. We ignore the fine grained distinctions of meaning between sentences and phrases, introduced due to the speaker’s evaluation of the situation, connotation of the terms used, change of modality, etc. For example, consider sentences (3) and (4).

(3) The school said that their buses *seat* 40 students each.
(4) The school said that their buses *cram in* 40 students each.

Here, “seat” and “cram in” are not synonymous: they carry different evaluations by the speaker about the same situation. We however consider sentences (3) and (4) to be (quasi) paraphrases. Similarly, consider sentences (5) and (6).

(5) The school *said* that their buses *seat* 40 students each.
(6) The school is *saying* that their buses *might accommodate* 40 students each.

Here, “said” and “is saying” have different tenses. Also, “might accommodate” and “seat” are not synonymous, due to the modal verb “might”. We however consider sentences (5) and (6) to be quasi-paraphrases.

Note that this paper focuses on defining quasi-paraphrases. It does not provide direct implementation/application results of using them. We believe, however, that this work will allow computation-oriented researchers, to focus their future work more effectively on a subset of paraphrase types, without concern for missing important material and it will provide linguistics-oriented researchers, with a blueprint of the overall distribution of the types of paraphrase.

### 2. Paraphrasing Phenomenon Classified

Although approximate equivalence is hard to characterize, it is not a completely unstructured phenomenon. By studying various existing paraphrase theories - Mel’cuk (2012), Harris (1981), Honeck (1971) - and through an analysis of paraphrases, obtained from two different corpora, we have discovered that one can identify a set of 25 classes of quasi-paraphrases, with each class having its own specific way of relaxing the requirement of strict semantic equivalence. In this section, we define and describe these classes.

The classes described here categorize quasi-paraphrases from the *lexical* perspective. The lexical perspective defines paraphrases in terms of the kinds of lexical changes that can take place in a sentence/phrase resulting in the generation of its paraphrases.

1. **Synonym substitution**: Replacing a word/phrase by a synonymous word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This category covers the special case of genitives, where the clitic “’s” is replaced by other genitive indicators like “of”, “of the”, etc. This category also covers near-synonymy, that is, it allows for changes in evaluation, connotation, etc., of words or phrases between paraphrases. **Example:**
   (a) Google *bought* YouTube. ⇔ Google *acquired* YouTube.
   (b) Chris is *slim*. ⇔ Chris is *slender*. ⇔ Chris is *skinny*.

2. **Antonym substitution**: Replacing a word/phrase by its antonym accompanied by
a negation or by negating some other word, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Example:
(a) Pat ate. ⇔ Pat did not starve.

3. **Converse substitution**: Replacing a word/phrase with its converse and inverting the relationship between the constituents of a sentence/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase, presenting the situation from the converse perspective. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. Example:
(a) Google bought YouTube. ⇔ YouTube was sold to Google.

4. **Change of voice**: Changing a verb from its active to passive form and vice versa results in a paraphrase of the original sentence/phrase. This change may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. This often generates the most strictly meaning-preserving paraphrase. Example:
(a) Pat loves Chris. ⇔ Chris is loved by Pat.

5. **Change of person**: Changing the grammatical person of a referenced object results in a paraphrase of the original sentence/phrase. This change may be accompanied by the addition/deletion of appropriate function words. Example:
(a) Pat said “I like football”. ⇔ Pat said that he liked football.

6. **Pronoun/“Referenced noun” substitution**: Replacing a pronoun by the noun it refers to results in a paraphrase of the original sentence/phrase. This also often generates the most strictly meaning-preserving paraphrase. Example:
(a) Pat likes Chris, because she is smart. ⇔ Pat likes Chris, because Chris is smart.

7. **Repetition/Ellipsis**: Ellipsis or elliptical construction results in a paraphrase of the original sentence/phrase. Similarly, this often generates the most strictly meaning-preserving paraphrase. Example:
(a) Pat can run fast and Chris can run fast, too. ⇔ Pat can run fast and Chris can too.

8. **Function word variations**: Changing the function words in a sentence/phrase without affecting its semantics, in the appropriate context, results in a paraphrase of the original sentence/phrase. This can involve replacing a light verb by another light verb, replacing a light verb by copula, replacing certain prepositions with other preposition, replacing a determiner by another determiner, replacing a determiner by a preposition and vice versa, and addition/removal of a preposition and/or a determiner. Example:
(a) Results of the competition have been declared. ⇔ Results for the competition have been declared.
(b) Pat showed a nice demo. ⇔ Pat's demo was nice.

9. **Actor/Action substitution**: Replacing the name of an action by a word/phrase denoting the person doing the action (actor) and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Example:
(a) I dislike rash drivers. ⇔ I dislike rash driving.

10. **Verb/“Semantic-role noun” substitution**: Replacing a verb by a noun corresponding to the agent of the action or the patient of the action or the instrument used for the action or the medium used for the action, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. Example:
(a) Pat teaches Chris. ⇔ Pat is Chris’s teacher.
(b) Pat teaches Chris. ⇔ Chris is Pat's student.
(c) Pat tiled his bathroom floor. ⇔ Pat installed tiles on his bathroom floor.

11. **Manipulator/Device substitution:** Replacing the name of a device by a word/phrase denoting the person using the device (manipulator) and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Example:
   (a) The pilot took off despite the stormy weather. ⇔ The plane took off despite the stormy weather.

12. **General/Specific substitution:** Replacing a word/phrase by a more general or more specific word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Hypernym/hyponym substitution is a part of this category. This often generates a quasi-paraphrase. Example:
   (a) I dislike rash drivers. ⇔ I dislike rash motorists.
   (b) Pat is flying in this weekend. ⇔ Pat is flying in this Saturday.

13. **Metaphor substitution:** Replacing a noun by its standard metaphorical use and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Example:
   (a) I had to drive through fog today. ⇔ I had to drive through a wall of fog today.
   (b) Immigrants have used this network to send cash. ⇔ Immigrants have used this network to send stashes of cash.

14. **Part/Whole substitution:** Replacing a part by its corresponding whole and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words. Example:
   (a) American airplanes pounded the Taliban defences. ⇔ American airforce pounded the Taliban defences.

15. **Verb/Noun conversion:** Replacing a verb by its corresponding nominalized noun form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. Example:
   (a) The police interrogated the suspects. ⇔ The police subjected the suspects to an interrogation.
   (b) The virus spread over two weeks. ⇔ Two weeks saw a spreading of the virus.

16. **Verb/Adjective conversion:** Replacing a verb by the corresponding adjective form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. Example:
   (a) Pat loves Chris. ⇔ Chris is lovable to Pat.

17. **Verb/Adverb conversion:** Replacing a verb by its corresponding adverb form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. Example:
   (a) Pat boasted about his work. ⇔ Pat spoke boastfully about his work.

18. **Noun/Adjective conversion:** Replacing a verb by its corresponding adjective form and vice versa, in the appropriate context, results in a paraphrase of the original
sentence/phrase. This substitution may be accompanied by the addition/deletion of
appropriate function words and sentence restructuring. *Example:*
(a) I’ll fly by the *end* of June. ⇔ I’ll fly *late* June.

19. **Verb-preposition/Noun substitution:** Replacing a verb and a preposition denoting
location by a noun denoting the location and vice versa, in the appropriate context,
results in a paraphrase of the original sentence/phrase. This substitution may be
accompanied by the addition/deletion of appropriate function words and sentence
restructuring. *Example:*
(a) The finalists will *play in* the Giants stadium. ⇔ The Giants stadium will be the
*playground* for the finalists.

20. **Change of tense:** Changing the tense of a verb, in the appropriate context, results
in a paraphrase of the original sentence/phrase. This change may be accompanied
by the addition/deletion of appropriate function words. This often generates a quasi-
paraphrase, although it might be semantically less accurate than many other quasi-
paraphrases. *Example:*
(a) Pat *loved* Chris. ⇔ Pat *loves* Chris.

21. **Change of aspect:** Changing the aspect of a verb, in the appropriate context, results
in a paraphrase of the original sentence/phrase. This change may be accompanied by
the addition/deletion of appropriate function words. *Example:*
(a) Pat is *flying* in today. ⇔ Pat *flies* in today.

22. **Change of modality:** Addition/deletion of a modal or substitution of one modal
by another, in the appropriate context, results in a paraphrase of the original sentence/phrase. This change may be accompanied by the addition/deletion of appropriate function words. This often generates a quasi-paraphrase, although it might be semantically less accurate than many other quasi-paraphrases. *Example:*
(a) Google *must buy* YouTube. ⇔ Google *bought* YouTube.
(b) The government *wants* to boost the economy. ⇔ The government *hopes* to boost the economy.

23. **Semantic implication:** Replacing a word/phrase denoting an action, event etc. by
a word/phrase denoting its possible future effect, in the appropriate context, results
in a paraphrase of the original sentence/phrase. This may be accompanied by the
addition/deletion of appropriate function words and sentence restructuring. This often
generates a quasi-paraphrase. *Example:*
(a) Google is *in talks to buy* YouTube. ⇔ Google *bought* YouTube.
(b) The Marines are *fighting* the terrorists. ⇔ The Marines are *eliminating* the terrorists.

24. **Approximate numerical equivalences:** Replacing a numerical expression (a
word/phrase denoting a number, often with a unit) by an approximately equivalent
numerical expression (even perhaps with change of unit), in the appropriate context,
results in a paraphrase of the original sentence/phrase. This often generates a quasi-
paraphrase. *Example:*
(a) At least 23 US soldiers were killed in Iraq last month. ⇔ About 25 US soldiers were
killed in Iraq last month.
(b) Disneyland is 32 miles from here. ⇔ Disneyland is around 30 minutes from here.

25. **External knowledge:** Replacing a word/phrase by another word/phrase based on
extra-linguistic (world) knowledge, in the appropriate context, results in a paraphrase
of the original sentence/phrase. This may be accompanied by the addition/deletion of
appropriate function words and sentence restructuring. This often generates a quasi-
paraphrase, although in some cases preserves meaning exactly. *Example:*

---

Bhagat, Hovy, and Hobbs

What is a paraphrase?

---

5
(a) We must work hard to win this election. ⇔ The Democrats must work hard to win this election.
(b) The government declared victory in Iraq. ⇔ Bush declared victory in Iraq.

3. Analysis of Paraphrases

In Section 2, we presented a list of lexical changes that define quasi-paraphrases. In this section, we seek to validate the scope and accuracy of this list. Our analysis employs two criteria:
1. **Distribution**: What is the distribution of each of these lexical changes in a paraphrase corpus?
2. **Human judgement**: If one uses each of the lexical changes, on applicable sentences, how often do each of these changes generate acceptable quasi-paraphrases?

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>% Distribution MTC</th>
<th>% Distribution MSR</th>
<th>% Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Synonym substitution</td>
<td>37%</td>
<td>19%</td>
<td>95%</td>
</tr>
<tr>
<td>2</td>
<td>Antonym substitution</td>
<td>0%</td>
<td>0%</td>
<td>65%</td>
</tr>
<tr>
<td>3</td>
<td>Converse substitution</td>
<td>1%</td>
<td>0%</td>
<td>75%</td>
</tr>
<tr>
<td>4</td>
<td>Change of voice</td>
<td>1%</td>
<td>1%</td>
<td>85%</td>
</tr>
<tr>
<td>5</td>
<td>Change of person</td>
<td>0%</td>
<td>1%</td>
<td>80%</td>
</tr>
<tr>
<td>6</td>
<td>Pronoun/&quot;Referenced noun&quot; substitution</td>
<td>1%</td>
<td>1%</td>
<td>70%</td>
</tr>
<tr>
<td>7</td>
<td>Repetition/Ellipsis</td>
<td>4%</td>
<td>4%</td>
<td>100%</td>
</tr>
<tr>
<td>8</td>
<td>Function word variations</td>
<td>31%</td>
<td>30%</td>
<td>85%</td>
</tr>
<tr>
<td>9</td>
<td>Actor/Action substitution</td>
<td>0%</td>
<td>0%</td>
<td>75%</td>
</tr>
<tr>
<td>10</td>
<td>Verb/&quot;Semantic-role noun&quot; substitution</td>
<td>1%</td>
<td>0%</td>
<td>60%</td>
</tr>
<tr>
<td>11</td>
<td>Manipulator/Device substitution</td>
<td>0%</td>
<td>0%</td>
<td>30%</td>
</tr>
<tr>
<td>12</td>
<td>General/Specific substitution</td>
<td>4%</td>
<td>3%</td>
<td>80%</td>
</tr>
<tr>
<td>13</td>
<td>Metaphor substitution</td>
<td>0%</td>
<td>1%</td>
<td>60%</td>
</tr>
<tr>
<td>14</td>
<td>Part/Whole substitution</td>
<td>0%</td>
<td>0%</td>
<td>65%</td>
</tr>
<tr>
<td>15</td>
<td>Verb/Noun conversion</td>
<td>2%</td>
<td>3%</td>
<td>100%</td>
</tr>
<tr>
<td>16</td>
<td>Verb/Adjective conversion</td>
<td>1%</td>
<td>0%</td>
<td>55%</td>
</tr>
<tr>
<td>17</td>
<td>Verb/Adverb conversion</td>
<td>0%</td>
<td>0%</td>
<td>65%</td>
</tr>
<tr>
<td>18</td>
<td>Noun/Adjective conversion</td>
<td>0%</td>
<td>0%</td>
<td>80%</td>
</tr>
<tr>
<td>19</td>
<td>Verb-preposition/Noun substitution</td>
<td>0%</td>
<td>0%</td>
<td>65%</td>
</tr>
<tr>
<td>20</td>
<td>Change of tense</td>
<td>4%</td>
<td>1%</td>
<td>70%</td>
</tr>
<tr>
<td>21</td>
<td>Change of aspect</td>
<td>1%</td>
<td>0%</td>
<td>95%</td>
</tr>
<tr>
<td>22</td>
<td>Change of modality</td>
<td>1%</td>
<td>0%</td>
<td>80%</td>
</tr>
<tr>
<td>23</td>
<td>Semantic implication</td>
<td>1%</td>
<td>4%</td>
<td>70%</td>
</tr>
<tr>
<td>24</td>
<td>Approximate numerical equivalences</td>
<td>0%</td>
<td>2%</td>
<td>95%</td>
</tr>
<tr>
<td>25</td>
<td>External knowledge</td>
<td>6%</td>
<td>32%</td>
<td>95%</td>
</tr>
<tr>
<td>26</td>
<td>Unknown</td>
<td>0%</td>
<td>0%</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 1
Distribution and Precision of paraphrases. Distribution may not sum to 100% due to rounding.

3.1 Distribution

We used the following procedure to measure the distribution of the lexical changes:
1. We downloaded paraphrases from two publicly available data sets containing sentence-level paraphrases: the Multiple-Translations Corpus (MTC) (Huang, Graff, and Doddington 2002) and the Microsoft Research (MSR) paraphrase corpus (Dolan, Quirk, and Brockett 2004). The paraphrase pairs come with their equivalent parts man-
We selected 30 sentence-level paraphrase pairs from each of these corpora at random and extracted the corresponding aligned and unaligned phrases\(^1\). This resulted in 210 phrase pairs for the MTC corpus and 145 phrase pairs for the MSR corpus.

3. We labeled each of the phrase pairs with the appropriate lexical changes defined in Section 2. If any phrase pair could not be labeled by a lexical change from Section 2, we labeled it as unknown.

4. We finally calculated the distribution of each label (lexical change), over all the labels, for each corpus. Table 1 shows the percentage distribution of the lexical changes in the MTC (column 3) and MSR corpora (column 4).

### 3.2 Human Judgement

In this section, we explain the procedure we used to obtain the human judgement of the changes that define paraphrases from the lexical perspective:

1. We randomly selected two words or phrases from publicly available resources (depending on the lexical change) for each of the lexical operations from Section 2 (except external knowledge). For example, to obtain words for synonym substitution, we used WordNet (Fellbaum 1998) (and selected a word say “buy”); to obtain implication rules for semantic implication, we used the DIRT resource (Lin and Pantel 2001); etc.

2. For each selected word or phrase, we obtained five random sentences from the giga-word corpus. These sentences were manually checked to make sure that they contain the intended sense of the word or phrase. This gave us a total of 10 sentences for each phenomenon. For example, for the word “buy”, one of the selected sentences might be:

   (a) They want to buy a house.

3. For each sentence selected in step 2, we applied the corresponding lexical changes to the word or phrase selected in step 1 to generate a potential paraphrase\(^2\). For example, we might apply synonym substitution to sentence (a) above and replace the word “buy” with its WordNet synonym “purchase”. This will result in the following sentence:

   (b) They want to purchase a house.

4. For the phenomenon of external knowledge, we randomly sampled a total of 10 sentence pairs from the MTC and MSR corpora, such that the pairs were paraphrases based on external knowledge.

5. We gave the sentence pairs to two annotators and asked them to annotate them as either paraphrases or non-paraphrases. For example, the annotator might be given the sentence pair (a) and (b) above and she/he might annotate this pair as paraphrases.

6. We used the annotations from each of the annotators to calculate the precision percentage for each lexical change. The final precision score was calculated as the average of the precision scores obtained from the two annotations. Table 1 shows the percentage precision (column 5) of lexical changes in this test corpus.

7. We finally calculated the kappa-statistic (Siegal and Castellan Jr. 1988) to measure the inter-annotator agreement. A kappa score of $\kappa = 0.66$ was obtained on the annotation task.

---

1 We assume that any unaligned phrase is paired with a null phrase and we discard it prior to the analysis.

2 The words in the new sentence were allowed to be reordered (permuted) if needed and only function words (and no content words) were allowed to be added to the new sentence.
4. Conclusion

A definition of what phenomena constitute paraphrases and what do not has been a problem in the past. While some people have used a very narrow interpretation of paraphrases—paraphrases must be exactly logically equivalent—others have taken broader perspectives that consider even semantic implications to be acceptable paraphrases. To the best of our knowledge, outside of specific language interpretation frameworks (like Meaning Text Theory (Mel’cuk 1996)), no one has tried to create a general, exhaustive list of the transformations that define paraphrases. In this paper we provide such a list. We have also tried to empirically quantify the distribution and accuracy of the list. It is notable that certain types of quasi-paraphrases dominate while others are very rare. However, it is also observed that the dominating transformations vary based on the type of paraphrase-corpus used, thus indicating the variety of behavior exhibited by the paraphrases. Based on the large variety of possible transformations that can generate paraphrases, its seems likely that the kinds of paraphrases that are deemed useful would depend on the application at hand. This might motivate the creation of application-specific lists of the kinds of allowable paraphrases and the development of automatic methods to distinguish the different kinds of paraphrases.

Acknowledgments

The authors wish to thank Jerry Hobbs for valuable comments and feedback.

References


