[*Translating and the Computer* 29, November 2007]

In other words: Using paraphrases in translation*

Pernilla Danielsson
University of Birmingham

Keywords: paraphrases, corpora, translation aids, meaning

1. Introduction

This paper is inspired by an event that took place 15 years ago. Recently graduated, I had eagerly started up a small consultancy company in Sweden with the aim of producing new translation software. The first collaborative project I was involved in was to offer consultancy around a translation memory called Eurolang Optimizer, where my task would involve producing Swedish localisation. The translation aid, as all other translation memories on the market, had as its main feature a database where it would store previous translations and later offer them as suggestions when the same source sentence was found in subsequent translation work.

With only a broken demo to show for my work, I was invited with my colleagues to the translation department at the Swedish Foreign Ministry to demonstrate this amazing new tool. I had carefully rehearsed every mouse click to avoid any bugs and I managed to get through the pitch without any catastrophes. To my own recollection, I had done a good job.

Following the presentation, only one question was raised: 'Given that politicians would not like to be seen to say the same things twice, and that our task as a translator is more often to rephrase and paraphrase rather than repeat, how do you envisage we would use this tool?'

I walked away without a sale. But the question of paraphrases in translation has stuck with me for all these years; how would one find a way to study authentic paraphrasing to be used for translations? This paper is devoted to the study of naturally occurring paraphrases in translations. The presentation is in four parts; firstly, a discussion about what is understood to be a paraphrase. Secondly, a corpus consisting of multiple English translations of the same source texts is presented as the language resource used in this study (The Transfer Corpus). Thirdly, a study focusing on the identification of

* [This paper was published in the conference proceedings in a draft form. It has been cleaned up by the editor of Machine Translation Archive.]

paraphrases in this corpus is presented and the findings categorised. These findings are marked up and stored in a database. The concluding remarks discussed the potentials of using this databases as complement to current translation aids.

2. Naturally occurring paraphrases in text

Paraphrases hold important information about how meaning is created in texts, yet little work has been done exploiting this resource in linguistic. Instead, the most active area of paraphrase research is in Machine Translation. Being able to automatically identify and mark-up authentic paraphrases in text could provide a most useful addition to any translator's aids. Much of the current research focuses on using paraphrases to evaluate MT output, see for example Zhou et al (2006) and Owczarzak et al (2006).

Translation has often been referred to as the art of paraphrasing a text from one language into another. As there is always more than one way to phrase a statement, then the result in the target text depends on the translator's choice. Finding alternative ways to word an utterance could be a useful tool for any creative language user, not just the translator, but few writers may be under as much influence whilst creating a text as the translator; source text wording as well as contextual settings are known to often affect the choices made by translators. Having a tool that could suggest alternative authentic phrasing would offer valuable input in the translation process. Although there is an emerging field of research on paraphrases, there is still very little said about *how* we paraphrase. Instead, in most publications on the subject, the goal is reached once the paraphrase had been identified (and is later put into good use for example for Machine Translation as mentioned above).This study should be viewed as a first attempt to study and categorise paraphrases and, although a translation tool is alluded to, the research is not yet ripe for such an implementation. In the extension to this project, we envisage a tool that will offer alternative words, phrases or even grammar to that of the translator's first choice.

The emphasis in this study here will be on identifying *naturally occurring* paraphrases, which should be distinguished from other types of paraphrases. This study does not primarily take an interest in the data found in thesauri, nor does it include the usual teaching of paraphrasing in language education. The main reason for this is that whereas most thesauri offer a list of near-synonyms, they tend to be single word items. In the case of paraphrases used in language learning, they tend to be more of a simple reforming of the same expression ('the pen is on the table, on the table is the pen') and situations are

made up in which students have to paraphrase statements which do not necessarily form a naturally sounding unit.

Naturally occurring paraphrases embrace a wider category. When a source is paraphrased, the wording is changed and the meaning is both ‘the same’ and ‘different’. Which is often exactly what is needed in translation; it is not that a translator cannot come up with a translation; it is more often an urge to find a better sounding way of expressing the statement.

A simple description of paraphrase is:

> *A paraphrase expresses a statement, a phrase or a single word, in some other words.*

Pparaphrases are used in our language for many reasons: they are there to clarify, explain, describe, define, transfer and/or reformulate an expression and, as such, they are vital for exploring natural language semantics. In short, paraphrase is used when an aspect of meaning is contentious or doubtful.

The problem is that they are difficult to identify in naturally occurring text. Because a paraphrase, by definition, does not consist of the same string of items as the original (paraphrases of *table,* usually do not include the string ‘table’), it is not possible to search a corpus of naturally occurring texts for paraphrases. To annotate a corpus so that paraphrases are tagged and are retrievable involves not only manual effort but also a large degree of subjective judgement. It is true that paraphrases are sometimes signalled in a text. Phrases such as *in other words, is the technical term for* or even the straightforward *means* often indicate the presence of paraphrase, but there is no requirement that a paraphrase should be so signalled. Identifying paraphrases within and, especially, between texts is no easy matter and, there being no guidelines as to what does and does not constitute paraphrase, identification is often a matter of intuition. There is a clear need, therefore, for studies which start from a non-intuitive identification of paraphrase.

Current translation tools in the form of MT products or translation memories usually only offer one possible translation. In fact, it has been highlighted as a problem in MT that systems repeat the same monotonous way of phrasing an utterance over and over again . An obvious solution to this would be to include paraphrases into already existing translation tools, however, linguistics has so far failed to offer a suitable description of the production of paraphrases.

3. Using language corpora to identify paraphrases

A useful starting-point for the study of paraphrase is a set of texts that can be identified as paraphrases of each other on external rather than internal criteria. Here, texts that have more than one translation into English are compiled into a corpus, the Transfer Corpus. Similar corpora used in paraphrase research have been reported by Barzilay & McKeown (2001) and Iordanskaja et al (1991). Other types of corpora that have been used looks at paraphrases in summarised texts, i.e. where the shortened phrase is aligned with its counterpart in the longer text (Zhou et al 2006). These types of paraphrases differ from the one in this study as they also have a goal to be short. Another type of corpus that may be used consists of revised versions of texts. Research on this type of data has been carried out by Falvey (1993), Utka, (2004) and John (2005), although they have not been focusing on paraphrases per se, but instead the revisions. Revisions offers a slightly alternative view on paraphrases as it also involves an evaluative feature, the latest revision is considered the best version. A fourth type of corpus that may be used for paraphrasal studies are the more traditional parallel corpora, consisting of source texts aligned with their target texts, again these have been used in MT research (Callison-Burch et al 2006) by searching for one and the same phrase in one language and assuming that the corresponding segments in the other language are paraphrases.

The Transfer Corpus consists of several translated English versions of the same original text. The main obstacle for compiling a language resource like this is that only very few texts are translated into English several times. This imposes heavy restrictions on the available resources. In this study the chosen text is Plato's *Republic*, which has been the object of numerous translations into English. The project as a whole also includes other texts with multiple translations such as Selma Lagerlöf's 'Gösta Berling's Saga' and Dante's 'The Divine Comedy'.

Many scholars have argued that translation should be viewed as a way of transferring texts from one language by *paraphrasing* it into another. As such, each translated text can be seen as a paraphrase of the original, and two translated texts can be seen as paraphrases of each other, or to have a paraphrastic relation.

The fact that the Transfer corpus consists of parallel texts allows it to make use of existing techniques for sentence aligning text (Danielsson and Ridings 1997). Each pair of aligned sentences may then be regarded as paraphrases of each other, based on external criteria alone. In this way the identification of

paraphrases can be done without recourse to intuition or to extensive manual processing.

In this study, the parts of the translated texts that are identical are ignored. Instead_the focus lies on what will be referred to as differential paraphrases, where the texts are different from each other. The example 1 below is an illustration of two aligned segments from the corpus that can be said to hold a paraphrastic relationship, i.e. can be said to be paraphrases of each other.

> (a) And about knowledge and ignorance in general: see whether you think that any man who has knowledge ever would wish to have the choice of saying or doing more than another man who has knowledge. Would he not rather say or do the same as his like in the same case?
>
> (b) In any branch of knowledge or ignorance, do you think that a knowledgeable person would intentionally try to outdo other knowledgeable people or say something better or different than they do, rather than doing or saying the very same thing as those like him?

Ex. 1. A paraphrastic relationship between two sentences.

4. Identifying paraphrases in naturally occurring text

The methodology of the study replicates that used in most corpus studies of translation. Initially, a search is conducted on the verbs *see*, *saw* and *seen*. In order to be captured in the search, a sentence in one of the texts will include a form of one of these verbs; the equivalent sentences in the parallel texts may or may not include the same verb. Sentences that are not formally identical are then identified as in a paraphrastic relationship. The degree of difference may vary, as illustrated in examples 2 and 3 below.

> (a) ***'everyone saw** that...'*
> (b) ***'it was clear to all** that...'*

Ex. 2 Illustration of the paraphrases *'everyone saw'* and *'it was clear that'*.

Example 2 illustrates a well-known concept in translation studies, that what is equivalent in translation is not the word but the phrase or sentence. The equivalence here is not between the two words *saw* and *clear* but between the two phrases.

(a) *then I saw **what I had** never seen before*
(b) *then I saw **something I'd** never seen before*

Ex. 3 Illustrations of the paraphrases *'what'* and *'something'*.

Having performed a closer investigation of the corpus data, we found that the phrases lend themselves to a categorisation based on their types of differences. Three categories dominate our results:

(1) Alternative lexis
(2) Alternative phrases
(3) Alternative grammar

However, these categories do tend to overlap, and one and the same segment may be marked up as belonging to several categories. Example 4 is an illustration of 'alternative lexis' because the word *see* is present in (a) but not in (b). It illustrates alternative grammar because (b) is in interrogative mood whereas (a) consists of a declarative followed by an interrogative. It is also an example of alternative phrases because the phrase *see whether you think* in (a) could be replaced by *do you think* in (b). In addition, *any man who has knowledge* in (a) could be replaced by *a knowledgeable person* in (b), and so on through the example. The equivalence is not between one word and another but between a phrase and a phrase.

(a) And about knowledge and ignorance in general; see whether you think that any man who has knowledge ever would wish to have the choice of saving or doing more than another man who has knowledge. Would he not rather say or do the same as his like in the same case?

(b) In any branch of knowledge or ignorance, do you think that a knowledgeable person would intentionally try to outdo other knowledgeable people or say something better or different than they do, rather than doing or saying the very same thing as those like him?

Ex. 4 Example of segments with a paraphrastic relationship, especially concerning *'see'* and *'do you think'*.

The following sections will be devoted to a closer look at the findings in each category.

4.1.Alternative Lexis

This is the simplest form of paraphrase; a word is exchanged for another word or a combination of a few words. This is the category that resembles the information captured in a thesaurus. Taking the outset from a single word, *the look up word*, the thesaurus will offer a selection of other single words, or in a few cases, a short phrase as possible alternatives.

Examples 5 and 6 show some of these Alternative Lexis items in context, where the word *see* is paraphrased into *look at*

> (a) He drew near, and they told him that he was to be the messenger who would carry the report of the other world to them, and they bade him hear and *see* all that was to be heard and seen in that place.
> (b) When Er himself came forward, they told him that he was to be a messenger to human beings about the things that were there, and that he was to listen to and *look at* everything in the place.

Ex. 5 Paraphrases in context; *see* and *look at.*

> (a) Let us rise soon after supper and *see* this festival; there will be a gathering of young men, and we will have a good talk.
> (b) After dinner, we'll go out to *look at* it. We'll be joined there by many of the young men, and we'll talk

Ex. 6 Paraphrases in context; another illustration of *see* and *look at.*

Clearly, this categorisation allows us only to address the surface level of what is actually going on in a paraphrased sentence as the one above. A more in-depth analysis of what has changed between the two sentences would need to include much more. In example 6 above, for instance, the role of the addresser alters between the two versions. In (a), the addresser is likely to be a participant in the festival while, in (b), the addresser is only an observer of the festival beside the surroundings.

Verb form/paraphrases	possible paraphrase	possible paraphrase
See	look at	determine
Saw	caught sight of	looked upon
Seen	appear	occur
Hear	listen to	listen
heard	mentioned	listened

Table 1: Some, findings belonging to the, category Alternative Lexis

Table 1 is used as an illustration of the findings that have fallen into this category.

4.2 Alternative Phrases

The category of Alternative Phrases takes into account the fact that the words in our study *see* and *hear* are often part of large units of meaning, i.e. multi-word units, such as *let's hear it.* A larger unit with a clear semantic value may very well still be paraphrased with only one word, however, in our data we find that two larger units paraphrase each other. The concept of replacement is important here. In an example of Alternative Lexis (e.g. example 9), a single word in one sentence (e.g. *see}* may replace a multi-word unit in another (e.g. *reach the study of).* In Alternative Phrases, to make sense of the sentences, more than one word would have to be involved in the replacement. Here are some examples of phrases which could replace each other:

HEAR	*when I hear you say that*	*even as you were speaking*
	let us hear	*continue to explain*
HEARD	*You have often heard me say*	*you know very well that I am going to say this*
SAW	*everyone saw that*	*it was dear to all that*
SEE	*But see the consequence*	*then, it follows... that*

Table 2. Examples of findings from the category Alternative Phrases

In example 7(a) below, *see* forms a meaningful unit with *but... the consequence.* Evidence for this as a unit comes from the paraphrase of the whole unit rather than of the individual item see: *it follows that* (7b). As before, the paraphrases have other consequences. Example 7(a) is an imperative, implying effort, whereas 7(b) construes a natural sequence of events.

> (a) ***But see the consequence****:* Many a man who is ignorant of human nature has friends who are bad friends, and in that case he ought to do harm to them; and he has good enemies whom he ought to benefit: but, if so, we shall be saying the very opposite of that which we [...]
>
> (b) ***Then, it follows***, Polemarchus, ***that*** it is just for the many, who are mistaken in their judgment, to harm their friends, who are bad, and benefit their enemies, who are good.

...Ex. 7. Illustration of findings in the category Alternative Phrases.

4.3 Alternative Grammar

The category Alternative Grammar is used when the two sentences in the pair differ in terms of grammar. This includes realization or omission of optional elements, such as *that* in noun clauses or the object pronoun in verb phrases as illustrated below in example 8.

he asked...
he asked him...

Ex. 8. Optional object pronoun as paraphrase.

It also includes variation in tense, aspect, and voice as in:

SAW	*(and) we never saw (her)*	*(and) we didn't see (her)*
	what he saw before was an illusion	*what he'd seen before was incense...*
HEAR	*did you ever hear*	*have you ever heard*
	you might to hear (them)	*(these things) must also be heard*

Table 3. Findings in the category Alternative Grammar.

Comparing the phrases above with their usage in a modern English Corpus may further be of assistance for the translator. For example, the phrase *'did you hear that'* occurs in The Bank of English corpus (a 450 million corpus of present-day English) 52 times. Whereas the phrase *'have you ever heard'* occurs 197 times, almost four times as frequent.

The alternations between positive and negative statements is another phenomena that we count as Alternative Grammar:

(a) ***Do you see that*** there is a way in which you could make them all. yourself?

(b) ***Don't you see that*** there is a way in which you yourself could make all of them?

Ex. 9. Paraphrase shift between positive and negative statement.

5. Conclusion

This study gives an indication of the types of paraphrases that can be identified in texts. Having classified the types of difference between paraphrases, and identified all the paraphrases in our corpus involving one of the words *see* and *saw*, we are able to quantify the different types. Table 4 below shows the number of each type of paraphrase.

	See	Saw
Alternative lexis	30	2
Alternative phrases	23	2
Alternative grammar	31	2
Other	27	11
Total	111	17

Table 4. Quantitative result from the identification of paraphrases of the verbs *see* and *saw*.

These figures suggest that there is an even distribution between paraphrases in the three categories, which shows that traditional thesauri-type of information only suffice to offer a third of all the possibilities. This is perhaps not surprising, as it is known that good translators translate phrase-by-phrase or sentence-by-sentence rather than word-by-word. It also confirms findings in recent corpus linguistics that support the importance of phraseology, as opposed to separate concepts of lexis and grammar, to language. As differences between phrases might be said to include differences between lexis and grammar, our data shows how frequently these two aspects of language work together in expressing paraphrase. Our findings also suggest that our data is a good source for the identification and classification of paraphrases, with a substantial number of differences being identified of each type for each of the words investigated. With the methodology established and tested it now becomes possible to carry out more extensive investigations on a much larger number of words, using automated techniques.

Once a larger database has been established, the data can be converted into a translator's tool. The tool can thus offer alternative phrasing to any selected word or phrasing in a current document. Although the restrictions imposed on the transfer corpus in terms of textual criteria (the types of text available in multiple translations) may pose a problem for the final tool however, judging by the result presented above it will still be able to offer more than what is available in current similar tools (i.e. online thesauri and dictionaries).

We believe it will also offer valuable input to machine translation systems, as it would complement the MT system in an area where they are currently struggling, namely to be able to identify more than one way of expressing a phrase.

References

Barzilay, R & McKeown, K.R. (2001) Extracting Paraphrase from a Parallel Corpus. In: Proceeding of ACL 2001:50-57.

Callison-Burch, C. Koehn, P. and Osborne, M. (2006) Improved Statistical Machine Translation Using Paraphrases. In *Proceedings from the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 17-24. New York: ACL.

Danielsson, P. & Ridings, D. (1997) Practical Presentation of a "Vanilla" aligner. In Reyle, U. and Rohrer, C. (eds.) Presented at the TELRI Workshop on Alignment and Exploitation of Texts. Institute Jozef Stefan. Ljubljana.

Falvey, P (1993) Towards a description of corporate text revision. PhD Thesis, University of Birmingham,

Iordanskaja, L., Kittredge, R. and Polguère, A. (1991) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, chapter 11. Amsterdam: Kluwer Academic Publisher.

John, S P (2005) The Writing Process and Writer Identity: investigating the influence of revision on textual and linguistic features of writer identity in dissertations. Unpublished PhD thesis, University of Birmingham, Birmingham, United Kingdom.

Owczarzak, K, Groves, D. Van Genabith, J. and Way, A. (2006) Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 86-93. New York: ACL.

Utka, A. 2004. English-Lithuanian Phases of Translation Corpus: Compilation and Analysis. *In International Journal of Corpus Linguistics 9:2:195-224.*

Zhou, L. Lin Chin-Yew, Munteanu and Hovy, E. (2006). ParaEval: Using Paraphrases to Evaluate Summaries Automatically. *In Proceedings from the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 447-454. New York: ACL.